UK Research and Innovation

Improving the evaluation of youth engagement with STEM

Scoping study

Improving the evaluation of youth engagement with STEM

Scoping study

Dr Asimina Vergou

April 2022

Acknowledgements

This report has been compiled by Dr Asimina Vergou.

The author would like to thank all the experts that actively participated in the interviews, providing constructive, informed and detailed feedback contributing to exploring how to improve the evaluation of youth engagement with STEM.

Special thanks to the UKRI's Public Engagement team for shaping this study, their guidance and continuous support: Dr Marianne Shelton and Dr Cassy King.

Table of Contents

Continued on next page

Executive summary

Through desk-based research and expert interviews, this study set out to investigate how funders and policy makers could drive coordination and improvements in the evaluation of youth engagement with STEM. The study explored the current landscape of evaluating youth engagement with STEM, gaps and challenges and key learnings from the evaluation practice in other sectors and key initiatives. Between February and March 2022, 18 project and programme evaluation reports and papers were reviewed, approximately 40 academic papers were analysed and synthesised, and 14 experts were interviewed.

The current landscape of evaluating youth engagement with STEM in the UK

Some progress has been achieved in evaluating youth engagement with STEM in the UK in the last decade. There is an increased understanding of the evaluation's value, and evaluation has become a recurring theme and topic in conferences and convenings. Funders, research groups, and STEM networks have been advocating for improving evaluation quality and developing and using evaluation frameworks and Theories of Change (ToC). There is more awareness of the need for evaluating practice concerning diversity, equity and inclusion (DEI), and some organisations are already measuring aspects of DEI. More evaluation work is being published in journals. Although progress has been made, there are still persistent issues and whilst there are pockets of good practice, there is also poor evaluation quality. There are in-house staff dedicated to the evaluation in larger organisations and an adequate budget to commission evaluation to consultants. In smaller organisations, evaluation is often not embedded into budget lines, and staff may not have the in-house skills to conduct or even commission the evaluation. Often there is no time for practitioners to engage in the evaluation on top of their other responsibilities. There is a lack of perception of evaluation as a skilled job, a lack of understanding of quality standards, and an understanding that evaluation should be embedded into a project from the start. Many evaluations rely on self-report and short-term measurement of impact. Sharing evaluation findings has increased but is still generally considered poor. The utilisation of evaluation findings is variable, and evaluation often focuses on positively endorsing a project rather than genuine reflection. There is a disparate network and not a strong community of evaluators focusing on STEM engagement.

Key issues

Common outcomes and shared measures

As part of an effort to standardise evaluation in the UK, there have been several attempts by funders, networks and research groups to develop frameworks that map out outcomes of public engagement with STEM. There is a need to encourage a strategic approach to investment in youth engagement with STEM; this should include having a long-term plan for that work and having agreed on common outcomes and clear goals. Moreover, there is a debate regarding the value and use of shared measures. Shared measures allow for aggregation and comparison of data across project or institutional evaluations and can be used to build the evidence base and justify the need for further funding. However, there are concerns regarding the feasibility of agreeing on common outcomes and the validity of using standard methods to cover a wide range of engagement methods, designs, purposes and contexts.

Theory of Change (ToC)

It has been recommended that STEM engagement initiatives use a ToC approach to critically examine whether what they deliver is likely to achieve the intended aims. There are already examples of STEM initiatives developing a ToC during the intervention design or as part of the evaluation. However, it is unclear what percentage of the STEM stakeholders overall are using ToC and whether they are using these to improve their practice.

Longitudinal evaluation

The evidence base for the long-term impact of youth engagement with STEM remains limited in the UK due to insufficient budgets, methodological challenges and the absence of accepted indicators or measures. There are only a few examples of longitudinal STEM engagement evaluations, such as for programmes running over the years that demonstrate the funder's and delivery organisation's commitment to using evaluation to improve the programme and evidence longer-term impact.

Methodological and technical issues

Evaluation designs often focus on assessing the impact at the end of a STEM intervention. There is a need to use evaluation more to inform the project development at the start and during its implementation. There has been an increase in the use of RCTs. However, critics argue that they import a medical model into education, which is problematic for looking at the impact on informal science learning experiences. There is a predominance of using interviews and surveys that rely on self-reporting, and there is a need to develop more direct and unobtrusive methods.

Measuring Diversity, Equity and Inclusion (DEI)

There has been some progress in measuring DEI in STEM engagement with the provision of guidance and tools such as The Equity compass. Evaluation has a role to play in supporting more equity-focused efforts by privileging the voices and lived experiences of non-dominant groups of young people, engaging young people in identifying desired outcomes as part of front-end evaluation, and ensuring that instruments and measures used are valid for the culture and context in which the evaluation is situated.

Need to change evaluation culture

Issues concerning evaluation culture include a lack of sharing evaluation learnings, including failures and being evidence-based. The STEM engagement sector is increasingly becoming more open to sharing evaluation findings and reports, but they lack the mechanisms and infrastructure. It is also important findings across evaluations are collated and synthesised in a format that is user-friendly and accessible. Evaluation is often conducted for accountability purposes, and income uncertainty deters grantees from sharing failures and what didn't work during a project. Practices in the STEM sector are rarely conceptually informed or evidence-based, and it is often unclear whether the evaluation findings will be used in the future to improve intervention.

Challenges and opportunities for evaluation due to Covid 19

Due to the Covid 19 pandemic, the provision of youth engagement with STEM moved to digital; hence suddenly, there was a need for new effective evaluation methods to capture impact and develop evaluators' capacity to understand and measure digital engagement. The increased use of digital platforms such as zoom created more flexibility for getting feedback from participants that face access difficulties. Another emerging opportunity was the possibility of comparing online and face-to-face provision of programmes and exploring a hybrid model of activities in the future.

Recommendations

In the UK, the youth engagement with STEM sector is varied, affected by many factors and different funders. Different STEM groups are working separately towards improving evaluation. Bringing change in the STEM sector requires leadership and convening to unify the dispersed efforts. The following recommendations suggest where efforts could be made by funders and policy-makers to bring systemic change:

1. Establish collaboration between funders and policy-makers for the creation of joint funding programmes to improve the evaluation of youth engagement with STEM and alignment in project evaluation requirements between different programmes to reduce burden on grantees.

2. Encourage debate and discussion and bring consensus in the sector in relation to the purpose of engaging young people with STEM, creating a shared common outcomes framework and an overarching ToC.

3. Support the development and use of ToC at project and initiative levels through funding, creating resources, and sharing case studies.

4. Support the robust data collection to track change and impact in the STEM sector by funding an online hub that will include tools for measuring impact, training, functionality to upload and analyse data from multiple STEM stakeholders. The online hub should enable easy, user-friendly and free access to a range of shared measures.

5. Invest in long-term evaluations of key youth engagement STEM programmes and the set-up of a longitudinal study to track change in youth engagement with STEM over time across the sector.

6. Influence the evaluation culture by enabling and encouraging sharing of evaluations more openly through investing in an online repository and framing new funding calls appropriately to encourage the sector to use evaluation evidence to improve practice rather than accountability.

7. Launch a dedicated funding programme for practitioner-researcher/evaluator collaboration to pilot innovative methods to measure youth engagement with STEM.

8. Invest in the development and sharing of tools to effectively measure DEI in youth engagement with STEM and in programmes that train and support evaluators professionally from diverse backgrounds.

9. Fund the expansion or creation of new communities of practice and networks for improving evaluation practice, fund the development of evaluation resources based on the gaps and what the sector needs and invest in professional development in evaluation for evaluators and practitioners.

10. Fund the creation and dissemination of tools and resources to help professionals who engage young people with STEM self-evaluate their practice.

Bringing systemic improvement in evaluating youth engagement with STEM requires long-term commitment and significant convening and joining forces across the many stakeholders in the sector. There are two different approaches that funders and policy-makers could co-invest and partner to adopt to coordinate the sector effectively:

- a large field-driven funding programme that will call for proposals related to what the sector has identified as needed to improve the evaluation of youth engagement with STEM.

- the set-up of an independent centre on improving youth engagement with STEM evaluation

1. Introduction

In 2012, a review of Informal Science learning in the UK (Lloyd et al., 2012) found that considerable effort and investment had been directed towards evaluation across the sector, mostly focused on process issues and capturing immediate or short-term outcomes. The review also highlighted challenges to evaluation such as limited sharing of learning, resource constraints (time, staff, skills, funding), lack of evidence on long-term impact, technical issues such as lack of clarity on evaluation methods, and cultural issues such as lack of institutional support and understanding of the value of evaluation. As a follow-up action, the National Forum for Public Engagement with STEM was established, with ensuring more effective evaluation as one of their priorities.

In 2019 the National Co-ordinating Centre for Public Engagement (NCCPE) published a survey of staff and volunteers involved in public engagement with STEM that revealed that the sector continues to face the same issues towards evaluation, such as the need for longitudinal studies, the need for appropriate resourcing, greater consistency and standardising, sharing failures and sharing evaluations. More recently, UKRI focusing on youth engagement with STEM, organised an event with key stakeholders to look at areas where the sector needs further support. The event concluded that one of the key areas that needs improving and further support was evaluation.

Within this context, this study was commissioned by UKRI to review how funders and policy-makers could drive coordination and improvements in how the sector evaluates and considers the impact of youth engagement with STEM (ages 5–19). The Key research questions that guided the study were:

- What is the current landscape of evaluating and assessing the impact of youth engagement with STEM (ages 5–19)?
- Where are the gaps in relation to assessing the impact of youth engagement with STEM?
- Who needs to be involved and how in order to improve evaluating youth engagement with STEM?
- What are key learnings from other sectors that can be used to improve the evaluation and assessment of the impact of youth engagement with STEM?
- What are some key initiatives and interventions that funders and policy makers could explore/consider further to support evaluating youth engagement with STEM?

1.1. Scope

This study has focused on STEM engagement activities in and outside of school for young people aged 5–19 years old. The formal and informal STEM engagement field is broad and diverse and includes one-off or series of experiences and interventions taking place in schools, designed spaces (such as science and discovery centres, zoos and aquaria, science museums, makerspaces), community spaces (such as STEM clubs, community organisations, STEM networks), events (such as science festivals, pop up café evenings, hackathons), support schemes (such as STEM mentoring, placements, bursaries/awards) and everyday forms of engagement (such as TV, media, websites, social media, books, magazines) (see Godec and Archer, L. 2021).

Youth engagement work is not a single or straight-forward type of activity. It can range from activities, approaches and initiatives that involve young people as participants (in which something is done ***for*** or delivered ***to*** young people), through to highly participatory and democratic approaches and initiatives in which young people can take a lead role in planning, delivery and decision-making, often as part of action towards social change (i.e., the approach is doing something ***with*** young people).

The formal and informal STEM engagement sector includes a diverse range of formats and approaches. These span from interventions that seek to inform and deliver STEM ***to*** young people (e.g., in order to increase the numbers and diversity of young people progressing through the 'STEM pipeline'), through to those which seek to use STEM as a means towards more equitable outcomes for young people and which empower young people to reconfigure and repurpose STEM to serve the social good (see Godec and Archer, L. 2021).

2. Methodology

The study was carried out between February and March 2022 and included a desk-based review of key literature, reports and academic papers published in the last decade, and expert interviews. The list of keywords used for the desk-based research, the types of experts interviewed, and the interview guides are included in the appendix. In total, 18 project and programme evaluation reports and papers were reviewed, approximately 40 academic papers were analysed and synthesised, and 14 experts were interviewed.

The results of evaluations of youth engagement with STEM initiatives are either studies published in peer-reviewed journals or nonpeer-reviewed evaluation reports, also known as grey literature. In this study, both types of publications were considered. With few incentives for practitioners and evaluators in non-academic positions to publish in peer-reviewed journals (National Research Council [NRC], 2009), the grey literature is a viable option for sharing knowledge from evaluations. Such sharing is encouraged and even mandated within the community. For example, in the past, funders from the National Forum for Public Engagement with STEM have encouraged their grantees to publish their evaluation reports on Collective Memory. Google Scholar, Google search, evaluation reports databases, e.g. STEM Learning and specific journals such as International Journal of Science Education, Part B and Research for All were used to locate project and programme evaluation reports and papers. The sample of reports is not representative of all youth engagement with STEM evaluation reports but provides some insights and indications about some trends in reporting evaluation findings.

Another primary data source for this study was a set of qualitative interviews with experts in youth and public engagement with STEM, youth engagement, education, arts and research and evaluation. Some of the 14 experts in this purposive sample were highly cited in the research/evaluation literature; others were active in national professional networks or recommended by other experts. The interviewees included public and youth engagement practitioners with past or current leadership positions, network representatives, evaluators, researchers and funders.

The semi-structured interview questions were developed based on this study's overarching research questions. Interviews were conducted through zoom and lasted approximately 1 hour. All interviews were audio-recorded and transcribed. The interview responses were used to map the landscape of current practice in evaluating youth engagement with STEM and identify recommendations for how to support the sector to move forward. Sometimes, what is cited from the interviews is not necessarily a common theme across multiple interviews but rather critical examples or quotes that illuminate particular gaps in the field or strategies for addressing those gaps. As a final caveat, with only 14 purposively sampled interviews, it is not possible to claim that these perspectives are representative of the youth engagement with STEM field.

3. The current landscape of evaluating youth engagement with STEM

In the last decade, experts agreed that some progress had been achieved in evaluating youth engagement with STEM in the UK. There is an increased understanding of the evaluation's value, and evaluation has become a regular theme and topic in conferences and convenings. Stakeholders such as funders, museum practitioners and academics are coming together to discuss how to improve evaluation practice.

> *'I think there's been an understanding that evaluation is important and should be done in strategic ways that can lead to real value.'*

The 2012 Wellcome review of informal science learning (Lloyd et al., 2012) has been considered influential in identifying evaluation practice's strengths and weaknesses and calling for better coordination to improve evaluation and support the sector to become more evidence-based. Other influential work includes the UCL's ASPIRES longitudinal research studying young people's science and career aspirations (see Archer L. et al., 2020) and Wellcome's Science Education Tracker, a survey into young people's attitudes towards and experiences of science education and careers (see Hamlyn, et al., 2020). Notable developments in the sector include the founding of the Science Museum's Learning Academy, which delivers research-informed training and resources to the STEM sector and the setting up of collaborative PhD programmes between the NHM, London, Royal Botanic Gardens, Kew and King's College, London which placed doctoral researchers to study and inform the informal science activities of the institutions. In the STEM sector, there has been increased advocacy by funders, research groups, and STEM networks for improving the quality of evaluation and developing and using evaluation frameworks and Theories of Change (ToC). More organisations and projects are developing evaluation frameworks and creating ToC. The sector has also moved in relation to engaging more diverse audiences and becoming more inclusive, and there is more awareness of the need for evaluating practice concerning diversity, equity and inclusion (DEI).

There are examples of STEM organisations that have changed their evaluation practice and now help others to do so. Namely, EngineeringUK has improved its approach to evaluating its own initiatives and now encourages the sector to do more evaluation and more meaningful evaluation and share their findings and learnings. 200 organisations have signed up to the Tomorrow's Engineers Code of Practice and evaluation is one of the pledges they are committed to. More evaluation work is also being published in journals such as Research for All, including efforts to collaborate with young people to co-design evaluative approaches that can better meet the programme's needs and objectives. Experts have pointed out that the driving forces for improving evaluation practice are funders rather than the people delivering a project. For example, the Research Excellence Framework requires academics to evidence the impact of their work, including their public engagement activities. Depending on the funder, the direction or standards of the evaluation may be different. Corporate funders of youth engagement activities are more interested in the return on investment, whilst funders such as the Science and Technology Facilities Council (STFC), with a science focus, are looking for a reliable, valid measurement of impact. Since funders predominantly drive the need for evaluation, practitioners use it to provide evidence of the impact of their work rather than an opportunity to learn and improve their delivery.

> *'Part of it has been driven by how funders have been expecting people to show the impact and value of their work. And I think people have a real desire to evidence and evaluate what they're doing, predominantly to satisfy funders that they're meeting the outcomes that they're claiming. I think there's been less progress in people evaluating for the purpose of learning themselves about what's working and how to improve their work. And I think that that definitely could be improved.'*

Whilst ***'there is progress, there is a need to understand the gaps'*** and challenges more holistically. Experts interviewed for this study and more recent surveys conducted by NCCPE (Bultitude, Verbeke and Duncan, 2015; NCCPE, 2019) reveal that although progress has been made in evaluation, there are still persistent

issues and whilst there are pockets of good practice, there is also poor quality of evaluation. Issues concerning resourcing evaluation have been highlighted. In larger organisations such as the NHM, London and the Science Museum, there are in-house staff dedicated to evaluation. There is also budget adequate to commission evaluation to consultants when needed. Smaller sites don't have the resources or the skills. Although the importance of evaluation is recognised, it is often not embedded into budget lines, and staff may not have the skills in house to conduct or even commission the evaluation. As a result, the quality of evaluation briefs and the allocated costs are low. Budgets are low, and in addition, with budget cuts, evaluation costs are the first that are being minimised. Often there is no time for practitioners to engage in the evaluation on top of their other responsibilities. Moreover, there is a lack of perception of evaluation as a skilled job, a lack of understanding of quality standards, and a lack of understanding that evaluation should be embedded into a project from the start. That results in missing opportunities to collect good quality data from a project during its duration, and an external evaluator is often brought at the end. As far as it concerns quality, many evaluations rely on self-report and short-term. Expectations concerning evaluation may also be unrealistic in terms of expecting a one-off experience to create long-lasting impact on an individual's life and choices.

> *'[There are] expectations for measuring long-term impact (from funders, or perception that this is a requirement from funders) and expectations, e.g. that a one-off museum visit can influence a student's aspiration and intention to take up a STEM career...this cause and effect is unrealistic to measure.'*

While some museum experts in the interviews highlighted that the sector might be more open to sharing evaluation findings, a small scale 2015 NCCPE survey (Bultitude, Verbeke and Duncan, 2015) showed that sharing is generally considered poor. One of the barriers is the lack of a mechanism to share evaluation effectively. The utilisation of evaluation findings is considered to be variable at a project level, and whilst some organisations were reported to use such findings internally, such learning was considered almost non-existent across different organisations. An evaluation often results in a positive endorsement of a project and lacks genuine reflection; there is an over-reliance on summative approaches with less utilisation of front-end or formative methods, and there is no developed robust underpinning methodology. Some of the key challenges identified were assessing longer-term impacts and overcoming a perceived lack of willingness to share evaluation findings (especially negative aspects or perceived 'failures').

There is also still a divide between academics and practitioners. Regarding commissioning evaluation to external providers, there is a disparate network and not a strong community of evaluators with a focus on STEM engagement. There are loads of professionals conducting evaluation, *'lots of pockets of individual freelancers, all quite small providers, with the exception of some bigger consultancies'*.

Experts have described the US STEM evaluation field as more mature in comparison to the UK; however, some of the persistent challenges are shared between countries. Academic papers on informal science education (ISE) and interviews with experts highlight that progress in the US is linked to changes in evaluation expectations, use and culture. The development of shared outcomes, the increasing attention to the quality of evaluation tools, the support for a community of professionals involved in evaluation, and the rising demand from funders for ISE organisations to report on their outcomes have all changed the work being done. NSF (National Science Foundation) has helped the field think differently about evaluation by requiring evaluation as part of its funding requirements and by investing in capacity building among evaluators, practitioners, and researchers.

The establishment of informalscience.org, an NSF-funded website and comprehensive database of project evaluations, has improved access to evaluation reports and academic papers and facilitated learning and connecting across projects. Several initiatives have synthesized learnings from evaluations of informal education and brought together the sector to discuss issues and build evaluation capacity (see table 1 for some indicative NSF funded projects). NSF, the National Academy of Science and the NSF-funded Center for Advancement of Informal Science Education (CAISE) are some of the key players coordinating efforts to identify and discuss key issues related to ISE evaluation (Fu, Kannan and Shavelson, 2019a).

Table 1: NSF funded projects focused on improving the evaluation of STEM activities

NSF funded project title, Period, Budget, Project description
Dimensions of Success (DoS) Observation Tool **2010–2014, $1,085,482** This observation tool pinpoints twelve indicators of STEM program quality in out-of-school time and is widely used in the US.
STEM Evaluation Community Project **2016–2020, $611,600** Through stakeholder meetings and two studies, the project convened the STEM Evaluation Community to better understand how to increase the capacity of evaluators to produce high-quality evaluations of NSF STEM education and outreach programmes and projects.
Building Informal Science Education: Supporting Evaluation of Exhibitions and Programmes with an informalscience.org Research Network **2010–2015, $909,311** The project synthesised 520 evaluation reports from informalscience.org to understand what the field can learn from such a rich resource and share the collective evaluation findings in an accessible way.
From Common Measures to Measures in Common: A Convening to Enhance Measurement of Outcomes of Afterschool STEM Programs **2018–2020, $250,000** The project, culminating in a conference, identified intended outcomes and measurement tools of afterschool STEM programs and explored where further work is needed.
Roads Take: A Retrospective Study of Program Strategies and Long-term Impacts of Intensive, Multi-year, STEM Youth Programs, **2019–2023, $1,006,867** This project focuses on six long-standing STEM youth programs to identify and describe the impact on the lives of alumni as well as identify pathways from program strategies to long-term outcomes.

In the US, the publication of some key resources also influenced practice. Indicatively, in 2011, CAISE and the Visitor Studies Association produced the Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects (Bonney et al., 2011) to support principal investigators and others working on NSF and other grants, particularly with respect to how to collaborate with the project evaluators for the greatest impact. Previous to this, NSF funded experts from a range of ISE sectors, including youth and community programmes, to develop the Framework for Evaluating Impacts of Informal Science Education Projects (Allen et al., 2008).

Groups that support the evaluation of youth engagement with STEM, such as the Visitor Studies Association, have been expanding. New interest groups focusing on evaluating STEM engagement have emerged as part of larger networks such as the American Evaluation Association (AEA). Several sector-specific evaluation communities emerged in STEM education through a bottom-up approach. Experienced evaluators who work with stakeholders from within a specific sector use multisite evaluation approaches, form communities of practice, have identified shared measures, trained a range of evaluators, and used the experiences and results from these processes to promote understanding of the evaluation practice and the informal learning. Examples of such communities range from science festivals (e.g., EvalFest,) and science centres (e.g. Collaboration for Ongoing Visitor Experience Studies [COVES], to after school programs (e.g. National Girls Collaborative Project) (Allen and Peterman, 2019).

Challenges related to evaluation capacity in STEM engagement institutions persist. Few settings have evaluation departments or dedicated evaluation staff, and the majority, especially small organisations, have no evaluation staff and depend on external evaluators from universities or private firms. Evaluation capacity might be housed among the non-evaluation staff of a STEM engagement organisation who may conduct evaluation themselves or contract evaluators. That option has challenges as there is no consistent training and professional development in evaluation, and job descriptions, e.g. for museum professionals, do not require expertise in evaluation.

The next sections of this report explore in more detail key issues that need to be addressed to improve the evaluation of youth engagement with STEM in the UK, followed by recommendations.

4. Common outcomes and shared measures

4.1. Common outcomes framework

Over the last decade, there has been some progress in defining a set of key outcomes in STEM that reach beyond content knowledge and include constructs such as sparking interest and building identity. Outcomes frameworks developed in the US and Europe use concepts that are narrow enough to be operationalised and assessed but also broad enough to capture the full range of possible outcomes desired by educators and designers (Allen and Peterman, 2019). As part of an effort to standardise evaluation in the UK, there have been several attempts to develop and use frameworks that map out outcomes of public engagement with STEM more generally or focus on a particular STEM area. Some funders have developed frameworks to evaluate impact across their funding programmes within their public engagement strategy. STFC's framework for evaluating public engagement (STFC, 2017) has been highlighted as a good example as the funder provided the framework and supported their grantees to use it.

> *'One of the best examples of trying to evidence and evaluate STEM engagement is STFC's work.... I think that their approach to creating an overarching way of evaluating their work, very similar to the Museums, Libraries and Archives Generic Learning Outcomes framework, that seems to me to be quite a sophisticated approach... It's not perfect... but it's in the right direction, and then they help their people who gain their funding to evaluate against their framework and support capacity building around that, and that has worked very well.'*

Other frameworks are sector-specific to support a particular STEM community on how to measure the collective and long-term impact of STEM on educational and career choices, such as the EngineeringUK's impact framework (EngineeringUK, 2021). This framework is based on a psychology model for behaviour change, and the aspiration is for EngineeringUK to keep iterating it in the future. It provides a tool for the engineering engagement sector to articulate what they are trying to achieve with their intervention, assess it, and understand what the sector is collectively trying to achieve. EngineeringUK is delivering webinars and using other ways to disseminate it; however, feedback from the practitioners is that they find it challenging to use and apply it.

Science capital is a concept developed through the ASPIRES project to understand uneven patterns in science participation (Archer et al., 2015) and has been increasingly used as a framework to evaluate youth engagement with STEM (see DeWitt, 2019; Hope-Stone Research, 2018). Most of the experts interviewed referenced Louise Archer's work as influential in the sector. The science capital approach enabled practitioners to reflect on their work and change their practice.

> *'One of the good things that have happened is Louise Archer and her research team's Science capital work because I think that gives us a better understanding of what it is that we are doing and, therefore, what we might be contributing to... so I think that is a helpful thing. And I think that is also something that has a lot of traction within the practitioner community as well. So those people who are delivering STEM engagement activities understand, they can grasp what science capital is, and therefore they can see how their work plays into that. So, I think that's a helpful framework for us to then think, 'right, where does my work fit within that framework, and therefore, how might I evaluate the contribution that I'm making?'*

Specific projects were set up to support professionals to apply the concept in their contexts, such as The Science Capital in Practice programme (Science Museum Group, 2021). The two-year programme was a collaboration between the Science Museum Group and the UK Association for Science and Discovery Centres (ASDC) and engaged 15 UK science centres and museums aiming to establish a growing community of good practice around the application of science capital principles in informal science settings.

Some issues raised about the application of science capital is that its language is too academic, and some practitioners may find it difficult to relate to it and apply it in their practice. Characteristically, when it was used in an evaluation of a youth engagement project it was found that younger students couldn't respond to the relevant survey questions (see Shimwell et al., 2021). In addition, it was mentioned that in practice, the concept has been oversimplified and has become a 'new deficit model', measuring whether communities have high or low science capital. A US STEM expert questioned how much diversity there is around the science capital approach and whether there are different views of science capital.

> *'It's quite a complex framework. Professionals need to get into the theory behind it in order to understand how to use it properly. But I'm not sure to what extent people use it…it's quite a complex set of concepts to understand in order to put it into practice.'*

Another influential framework has been the Generic Learning Outcomes which was developed as a tool for museums, libraries and archives to demonstrate the outcomes and impact of users' learning experiences. The framework provided practitioners, government and funders with a meaningful way to describe and evidence the impact of museum experiences on visitors and report on these collectively. Its development was a high-profile project with many expectations. Initially, its development created anxiety as the museum sector didn't know what evaluation was at the time or who might be able to do it. But it has been adopted widely both within and beyond the UK, e.g. in Sweden and the USA. Its success lies not only in its wide adoption but also in establishing the idea in museums that it is crucial to have a framework to measure outcomes to demonstrate impact collectively.

Experts' views converged that there is a need to encourage a strategic approach to investment in youth engagement with STEM; this should include having a long-term plan for that work, having agreed common outcomes and clear goals, and having a clear approach to the sort of partnerships that might underpin it. However, before embarking on creating an outcomes framework, the STEM sector needs to make more explicit the different agendas, to agree and clearly articulate the purpose(s) of engaging young people with STEM and look at the impact from the learner's perspective rather than the intervention's perspective.

> *'You've got to know what it is that you're trying to do in order to evaluate something.'*

> *'It's always about what is the point of youth engagement with STEM…I don't think it is well defined…and because it's not well defined, it's very hard to evaluate. And I think we also tend to approach evaluation from the point of the intervention rather than the point of the learner or the participant who's involved. So we try and track the impact from…a visit to a science centre or having a scientist go into the classroom, or whatever it is…You can't track the impact from that because the young person's life experience is very noisy; you don't know what they've got before they come in, you don't know what they engage with afterwards. So we can't track it...it's not to say we shouldn't be doing long-term evaluation studies; it's just that we don't do them from the point of the interactive intervention.'*

One of the experts suggested that efforts should focus on front-end evaluation to ensure that the project design is based on effective approaches and then focus on evaluating impact by having an open mind to identify any potential outcome achieved and not narrowing it down to pre-determined ones.

Defining a set of common, agreed outcomes in the STEM sector is not an easy task. STEM education includes many different contexts (in and out of the school class), participants and practices and comparisons of outcomes across projects may be difficult. In addition, any attempt to create an overarching common framework should acknowledge and build on frameworks that stakeholders in the STEM engagement sector are already using.

> *'I think it's really hard to do. And I think that what you end up is potentially an outcomes framework that doesn't really capture the nuance of what's needed and change needed in the sector because it kind of becomes the… lowest common denominator. But what can we all agree on?'*

The importance of a funder driving the creation of a common outcomes framework with appropriate resources overtime was also pointed out.

> *'If a funder decided that they were going to create some kind of outcomes framework and that they were going to align their funding opportunities with realising certain outcomes, and then seeing how each of those things contribute to it, yeah, I would love it. And I think that that idea of then doing a longitudinal study over time where the effort of that longitudinal study was borne by the funder, not by the individual projects, which of course all stop funding their evaluation, the minute their funding stops. I think it's really tantalising. But … I think it will take a lot of time and a lot of energy and may take far more resource than anticipated.'*

4.2. Shared measures

In the US STEM sector, the debate has moved from whether a common outcomes framework is feasible to the value and use of shared measures. The intent is to build the capacity of evaluators to measure common outcomes of STEM engagement experiences. A shared measure is defined as an instrument developed to measure a particular outcome common across a range of programmes, projects, or the STEM field. The term "shared measures" focuses on creating or using rigorous measures that can be applied across programmes that address the same outcome. Although similar in meaning, the focus on "measurement of a common outcome" instead of a "common measure" is intentional. This stresses that evaluators need to shift focus from the measure to what any given instrument is measuring, how it relates to a programme's outcomes and the reliability and validity of the instrument. Several sector-specific instruments are developed in the US, predominantly observation tools and surveys (see case study on the Assessment Tools in Informal Science-ATIS) and open conversations have taken place about whether and how to use them. Training is crucial to enable evaluators to understand the reliability and validity of different instruments and make an informed choice and determine which one(s) might be a good fit for their projects while remembering that shared instruments are only part of a complete evaluation (Allen and Peterman, 2019; Grack Nelson et al., 2019). In the US, NSF supports convenings of the sector to discuss the need and process for developing shared measures and support the development of a variety of different approaches to measuring common outcomes, which allow to find out what works for whom in what context and why.

A major benefit of shared measures is that they support evaluators in conducting high-quality evaluations. When a shared measure is already available for an outcome of interest, evaluators may not need to develop an instrument from scratch, saving time and money. Moreover, using an instrument that has already been tested for an outcome in similar contexts to the current evaluation can increase evaluators' and clients' confidence in the quality of their evaluation data. Shared measures also allow for aggregation and comparison of data across project or institutional evaluations (Grack Nelson et al., 2019). Aggregation of data can help the field better understand the impact of STEM engagement experiences as well as identify ways to improve projects from not just one experience but multiple experiences. The use of shared measures in evaluation helps build not only the evidence base of engaging young people with STEM but also justification for securing further funding and policy support. This is hard to do with evaluation data from one project; when multiple projects demonstrate impact on shared measures, it strengthens the case for supporting youth engagement with STEM experiences (Noam and Shah, 2013).

On the other hand, a key concern related to shared measures is their misuse if evaluators lack an understanding of validity and reliability or lack the skills of judging or creating quality data collection instruments. There are also concerns that evaluators may use a shared measure as the "go-to" instrument because it is available even if it may not be the best match for the evaluation. In addition, while a shared measure may meet the needs of an evaluation, overdependence on any single measure is problematic (Grack Nelson et al., 2019).

Moreover, there are concerns regarding the validity of using standard methods and indicators to cover a wide range of engagement methods, designs purposes and contexts. Like the experiences themselves, the outcomes, especially of informal STEM education, are idiosyncratic, personalized, and unpredictable. Theoretically, if multiple programmes address the same outcome, then a single instrument could measure

that outcome across programmes; yet many in the informal learning community largely resisted pressures to standardize their instruments, with the concern that informal learning programmers will begin to design narrow experiences that "teach to the test" (Allen and Peterman, 2019). Some tensions remain unresolved, such as the gains and losses of encouraging the development and use of different measures vs the potential scientific and political power of replication and comparisons that shared measures allow (Bell, 2020).

Looking into the evaluation practice in the UK STEM engagement, professionals may often be measuring similar outcomes but using slightly different measurement tools. Therefore, there is a need to coordinate them so that they will agree on measuring the same factors using the same measures to allow aggregation and comparison of the data. This can create tensions as systems may already exist within individual institutions using their unique instruments over the years. Highlighting and illustrating the benefits of using shared instruments will be crucial to bringing more alignment.

Case study | Assessment Tools in Informal Science (ATIS)

Of particular importance has been the development and use of shared measures in the afterschool sector in the US. In 2008, the Noyce Foundation funded a study (Hussar et al., 2008) on the state of shared measures for evaluating afterschool STEM experiences. The study recommended the development of an online database of measurement tools, the development of a questions bank that should be used across afterschool STEM evaluations, and the creation of two shared measures (a survey and observation tool) for use in STEM afterschool programs. The website Assessment Tools in Informal Science (ATIS) was developed as a result of the study's recommendations. ATIS is a repository of measures for evaluating youth outcomes in Out of School Time settings. The study's recommendations also led to the development of two shared measures for the STEM afterschool field: a survey called The Common Instrument Suite and the Dimensions of Success observation tool. These were created by the PEAR Institute (initially based at Harvard University), which also provides training on how to use the tools. There is a cost to use the Common Instrument Suite, a customizable ten-item survey and the data are collected in a centralized database so The PEAR Institute can provide programmes with individualized reports as well as compare findings to other programmes across the country (Grack Nelson et al., 2019). The instrument has been widely used in after school STEM programmes for youth for many years; however, not everybody from the afterschool sector agrees with the methodology and the philosophy behind it, and other competing groups have developed different sets of instruments. Also, importantly, it was developed more than a decade ago; hence it isn't as current, especially as it doesn't cover measuring outcomes related to DEI.

5. Theory of change (ToC)

The value of a common outcomes framework, as an expert put it, lies in enabling the STEM sector to articulate a ToC that they can test against particular programmes and in enabling the STEM stakeholders to understand the 'enormity of what each one of them is trying to contribute to and recognise that. Archer, M. et al. (2021a) recommend that those who develop STEM engagement activities or programmes adopt a ToC approach to critically examine whether what they deliver is likely to achieve (or contribute towards) the intended aims.

A ToC is a predictive assumption about the relationship between desired changes and the actions that may bring about those changes. Putting it differently, *"If I do x, then I expect y to occur, and for these reasons"* (Connolly and Seymour, 2015).

ToC are being developed most often to evaluate existing initiatives, and they are also increasingly used to plan new initiatives. Some organisations are using ToC to revisit goals, assumptions, and activities of an existing initiative in order to understand and possibly change what they do (Taplin et al., 2013). One way to increase the chances that a change initiative will succeed is to explicate its ToC and then critically examine its reasoning about causes and effects (Connolly and Seymour, 2015). In the UK, the Arts Council England has been using ToC extensively as part of their new strategy (see case study). International development experts have cautioned that if ToC become a prescribed process as a condition of the funding, they may quickly become a compliance exercise and lose much of their value (Valters, 2015).

Case study | Arts Council England (ACE)

One of the successes in the approach of ACE is their use of ToC. Under their new strategy that runs from 2020-to 2030, every programme, every initiative, and every resource that they deploy will be backed up by a ToC, which will allow them to understand what the problem is, the rationale, and what does success look like through their particular intervention. They will then use survey tools and other mechanisms that they have developed to collect data, to better articulate the narrative around whether their investments are actually working or not. That has been a real success and a step-change in the way they have approached evaluation more recently. In the case of a collection of projects under certain programmes, they will develop a shared ToC in collaboration with other partners. They will then work with a supplier to try and understand through the ToC, what the common indicators are across those particular projects that build-up to the programme. And they will use these to look at progress and impact at programme level.

Several interviewees suggested that the STEM sector has progressed in the last decade in terms of using a ToC approach. Similar developments have been observed in the youth and arts sectors. It is unclear, however, what percentage of the STEM stakeholders are using ToC and also, importantly, whether they are using these to improve their practice. As part of this study, a review of youth engagement with STEM evaluation reports and papers, identified a small number of projects that had created and published their ToC (see PRiSE programme Archer, M. et al., 2021b; Crest awards, Husain et al., 2019; Generation STEM Work experience, Roy et al., 2021, The Outreach project, Davenport et al., 2021). For some of these initiatives, the ToC was developed during the design of the intervention, whilst others developed a ToC as part of the evaluation purposes. Projects are increasingly encouraged to clearly articulate the inputs, outputs, the steps in the process and the outcomes that they will aim to achieve. Experts interviewed suggested that the What Works Centres and especially the Education Endowment Foundation (EEF) have created a drive and encouraged the development of ToC for evaluating STEM projects; however, the influence from the EEF is linked to a small group of the

sector. Funders such as UKRI and Wellcome and research groups have been strong advocates for promoting and stimulating discussions around ToC. NCCPE's training has also been a positive influence in the STEM sector in terms of capacity building on developing and using ToC.

> *'We have found the idea of developing a ToC for an intervention, whether that's a small project or a larger kind of long-term initiative, a really helpful tool for teams to use to structure purposeful evaluation. We do quite a lot of work with people, just introducing them to the basics of that, a lot of practical support'.*

One of the interviewees explained that there is also some scepticism in the STEM sector around using ToC because they imply a linear progression and a cause-and-effect relationship.

> *'There are quite a few people who don't like ToC because of the implied cause and effect, this idea that if I do this, therefore, definitely this happens. And that linear trajectory, as if somehow, we have control over all of these things, which clearly, we don't. And so, particularly people who are more theoretically interested in the nature of change, kind of tend to step away from them. But we say that it's just a map to help you think about where to put your evaluative effort, not a map of reality, it's just, this is my best guess so far, and evaluation can tell me if my guess is even vaguely in the right place'.*

Others, have also cautioned that sometimes ToC tend to be far too linear and they risk 'squeezing out the space for learning (Mulgan, 2016). However, that depends on how they are used or misused and implemented. It has been suggested that organisations use ToC thinking which can create a framework to improve programme design, implementation, evaluation and learning (Valters, 2015).

The ToC approach is useful across the wider STEM engagement community. Using a ToC allows a community to identify individual activities that can be developed and aligned within the short- and medium-term outcomes yet still contribute to long-term outcomes. This should help increase their likely effectiveness rather than taking a purely short-term view. It also allows the identification of potential measures of success for the evaluation of activities (see case study of the ToC for the Outreach project). Indicatively, the ToC can be applied by those responsible for career guidance within school or college settings to ensure they are supporting the long-term outcomes, e.g. for the career aspirations and perceptions of young people. It has been suggested that solving the STEM skills shortage requires a number of nested ToC, each developed and shared by the key actors within the STEM space, e.g. companies, learned societies, government and charities. This would have two benefits: it would identify clearly how change is to be achieved, and it would allow a more holistic approach to encouraging young people to enter and remain in STEM careers. Working in partnership in this way is essential if the "wicked problem" of diversity in STEM is to be tackled effectively in a realistic fashion (Davenport et al., 2021).

One of the benefits of using a ToC process in STEM education and engagement is that it allows the development of solutions to complex problems such as the STEM skills shortage in a structured way (Davenport et al., 2021). ToC may shed light on causal mechanisms in youth engagement with STEM. These approaches require evaluators not only to ask whether a particular program produced the intended outcomes but also to examine why or why not (Fu, Kannan and Shavelson, 2019b).

In the youth sector, anecdotally, there has also been an increase in the organisations/ interventions which understand what a ToC is and use ToC. It is hard to generalise as it is difficult to identify a representative sample of the sector because 'nobody knows how big the sector is or what it includes'. However, overall, it is estimated that the percentages of organisations using a ToC are low.

One of the experts in the youth sector suggested that the quality of the ToC produced is low.

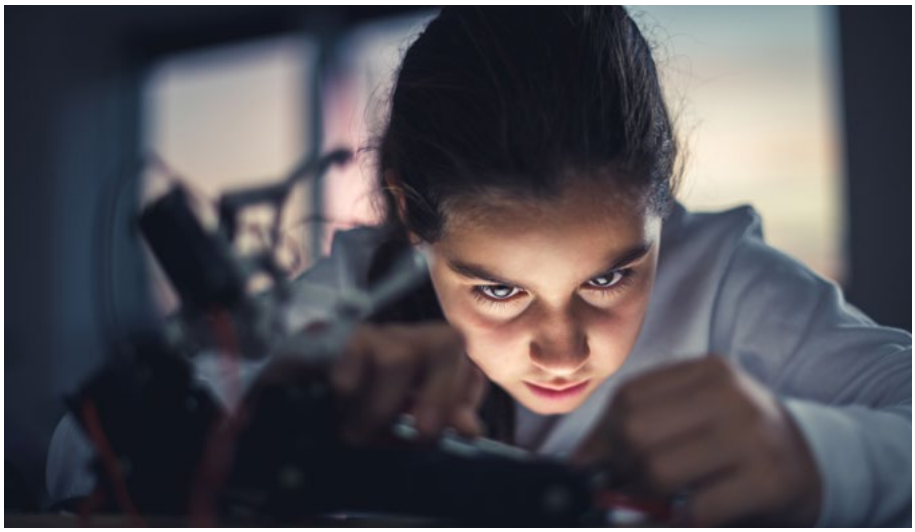
> *'ToC are very generic, very simplistic, and don't really allow for any kind of measurement or evaluation framework to hang off them. I also think that a lot of them aren't actually representative of reality. Most organisations don't really know what they are trying to achieve beyond anything generic'.*

Also, importantly if an organisation that has a ToC is asked to explain the rationale behind what they are delivering, they won't be able to justify it as an evidence-based decision. They may say *'it's because it's what they've always done because it's what the funder told them to do'.* Moreover, ToC may be increasingly produced, *'but they are not increasingly being used to actually make practice better'.*

Similarly, experts in the international development sector noted that ToC can often be 'superficial' and mechanistic and have not explored the deeper questions on assumptions. Instead, ToC should be considered as a flexible 'rough guide' that offers a unique set of perspectives, not a definitive, static prediction. A good ToC should be revised and adapted in the light of implementation experience, evaluation feedback and evidence (Valters, 2015). ToC approaches are growing in popularity, but they are often developed by evaluators. It is possible to create a ToC in a collaborative way which enhances the relevance and utility of evaluation for project teams and enables evaluation to become an embedded tool in their project development; this will require researchers/evaluators and practitioners investing time at the beginning of a project to develop through dialogue a shared ToC (Laing, 2022).

Case study | The Outreach project, Northumbria University STEM group

The Northumbria University STEM group (NUSTEM) set up an Outreach project in the North East of England with the aim to increase diversity and the number of young people choosing further study and a career in STEM. The project developed a ToC to shape child-focussed STEM interventions. The ToC identified how children, young people and their teachers and families could be engaged to increase the diversity and number of young people choosing STEM careers. The Outreach Project, which is a partnership of 10 organisations, worked with young people ages 2–19 years old across 30 schools. They provided ongoing interactions with children and young people, as well as their teachers and families. The Outreach Project was a multi-year intervention, with an intended long-term evaluation of children's qualification choices using the National Pupil Database (NPD) and planned to take place over a decade after the start of the project. A ToC approach was chosen to allow the evaluation to be clearly linked to the project's long-term aim through a chain of intermediate outcomes which are easier to track and evaluate. The ToC was developed through an iterative process. Backward mapping was used to clarify the steps required to achieve the overall aim of the project. The process involved identifying stakeholders, the changes required for each group of stakeholders (informed by research literature and professional expertise), categorising changes into short-, medium- and long-term outcomes, and identifying causal chains which linked the short-term with the long-term outcomes. The Outreach project team audited a number of the interventions that had already been delivered in schools against the draft ToC, looking at how well the fit was and whether the ToC could show the value of those interventions. Each outcome in the ToC was cross-referenced with research literature, and it was reviewed by experts. Developing a ToC for the Outreach project provided clarity when developing individual activities and offered a mechanism through which the desired outcomes could be made explicit for all the stakeholders. The ToC allowed the project team to recognise where their practice has strengths and limitations and encouraged a cycle of review and reflection to improve practice (Davenport et al., 2021).



6. Longitudinal evaluation

The need to increase investment in and conduct more longitudinal evaluation of youth engagement with STEM activities has been continuously articulated by the STEM engagement sector over the years, demonstrating that it is a persistent challenge the sector faces (Lloyd et al., 2012; Bultitude, Verbeke and Duncan, 2015; NCCPE, 2019). In addition, project evaluation reports in their recommendations suggest the need for further long-term evaluation to evidence whether the STEM activity achieved any long-term impact on the students (e.g. Thomas, Meakin and MacRae, 2022).

The evidence base for the long-term impact of youth engagement with STEM remains limited in the UK, especially in relation to demonstrating whether an increased uptake in STEM degrees in higher education occurs because of particular interventions (Archer M. et al., 2021a; Robinson and Salvestrini, 2020; Archer L. et al., 2014). Insufficient budgets, methodological challenges and the absence of accepted indicators or measures are all factors contributing to the lack of longitudinal evaluations (Lloyd et al., 2012). Practitioners report that they often struggle with long-term follow up, and even long-term running initiatives are rarely evaluated (NCCPE, 2019). The choice of appropriate methodological design is important when evaluators try to answer evaluations questions that might involve tracking changes over time and attribute long-term outcomes to the programme being evaluated (Fu et al., 2019c).

There are only a few examples of longitudinal STEM engagement evaluations, such as for programmes running over the years. These examples demonstrate the funder's and the delivery organisation's commitment to using evaluation to both improve the programme and to evidence any longer-term impact. For example, Nuffield Research Placements are evaluated over six years (see Cilauro and Paull, 2019). More recently, UKEngineering had planned to launch a longitudinal evaluation of the Big Bang Fair, which was disrupted due to Covid 19. When evaluating programmes that are running over many years, it is important for the evaluation budget, depending on the lifespan of the project, to link to a project iteration budget. If the evaluation indicates aspects of a programme are not working, there should be budget flexibility to redesign the delivery.

Experts pointed out the need for longitudinal studies to explore impact at a larger scale. The STEM sector in the UK is dominated by a large number of project evaluations, often conducted to satisfy the funders' requirements and *'not to deliver robust learning that then gets mobilised, nor is it building a kind of big collective picture'*. ***Instead,*** what is needed is *'coming up with an approach that allows a longitudinal kind of perspective on how to do this work well, how to evidence its impact and how to convince the government that it's working'*. Currently, there is no systematic way of telling the story of what the STEM sector is collectively achieving. The data doesn't aggregate in particularly meaningful ways, nor is it longitudinal. As a result, the STEM sector can't tell the story of what has been achieved over the years from the investment in engaging young people with STEM.

Another challenge in longitudinal evaluation is that for out-of-school interventions, there is no way to identify young people who participate and track them over several years in the same way that school interventions track the progress of young people, i.e. using the unique pupil number. For EEF funded projects, the data is archived in a repository, and it is possible to ascertain longer-term impact by tracking young people using their unique pupil numbers. EEF is currently looking into projects that initially focused on attainment in science and maths and exploring the longitudinal effects on post-16 destinations and choices. This would be possible by linking the NPD data to the longitudinal education outcome (LEO) database. Having the infrastructure to be able to track the progress of individuals engaged in STEM is important. Projects with a public engagement element that are funded through the Research councils are requested annually to track the impact of their grant through the researchfish database. Although the database is not very popular, it is an example of a mechanism that captures engagement data over time that might be valuable from a longitudinal point of view.

A lot of youth engagement with STEM activity is project-based, and as soon as the project finishes, the evaluation ends too. In addition, in many cases, it would be unrealistic for a funder to expect from a one-off experience, e.g. a museum visit, students to be influenced to take on a scientific career. Instead, it was

suggested funders invest in measuring the impact of longer-term interventions that are more likely to influence a young person's attitudes and choices significantly.

> *'I don't think that if you use a project-based funding model, you can expect the people that you fund to collect longitudinal data. I don't think it's acceptable. And I think it's very meaningless, particularly one-off interventions, to collect longitudinal data because there are so many other factors affecting the outcomes for the people that they engaged with the intervention…Capturing ten years after somebody went to the Science Museum if they're still interested in science is meaningless to me. But I think that for longer-term interventions, it's definitely worth it. And I think it's worth doing it at a collective level rather than an individual project level.'*

Funders have an important role to play in relation to investing in and influencing longitudinal evaluations, but their decisions are also based on the funding structures. For example, some trusts and foundations with long-term endowments are much more comfortable with investing in longitudinal evaluation rather than commissioners such as local authorities that have shorter budget cycles.

In the US, NSF is a major funder of youth engagement with STEM activity, research, and evaluation. Their awards are typically three to five years. Longitudinal evaluations are generally possible only through a series of awards, with no guarantee of securing these from start to finish as the process is competitive with each new funding cycle. One NSF-funded award, Roads Taken: A Retrospective Study of Program Strategies and Long-term Impacts of Intensive, Multi-year, STEM Youth Program, has taken a retrospective approach to studying youth engagement, including the impact of youth STEM programming on the lives of alumni as well as the pathways from program strategies to long-term outcomes.

7. Methodological and technical issues

Evaluation designs often focus on assessing whether the predetermined outcomes have been achieved or not at the end of a youth engagement with STEM intervention (e.g. Thomas, Meaking and MacRae, 2022; Shimwell et al., 2021; Gammon, 2012). There is a need to reconsider the focus of evaluation and look into how it could be used more to inform the project design at the start and during its implementation. When making critical decisions in a project, evaluation could be used to make evidence-based choices.

> *'I've always had this 80–20 model of evaluation. 80% of your funding for evaluation should be spent before and during the project design so that you're reliant on the most effective pedagogy that you know, and high quality and training for educators. And then at "the here", the 20%, smaller fraction, you should spend on trying to evaluate what impacts the world from the point of view of, 'we don't know what impacts them, maybe let's try and find that', rather than start off with a very narrow range of impacts and fish to see if they're there. So, I think the whole philosophy of evaluation of STEM should be more about getting it right in the first place rather than finding out too late that nothing's had an impact for…so many projects…What you want is to have people managing the project asking evaluation type questions so that they would say, 'this is a critical decision; have you considered all the evidence for making this decision [as part of] designing this education programme?"*

The issue of proportionality is also important to consider when making choices in designing the evaluation of an engagement activity. For example, for a large organisation that repeats a lot of programmes, it may be important to conduct an evaluation to look at effectiveness on a large scale. Also, when developing a new intervention engaging young people with STEM, it is crucial to evaluate it and look into whether it works, refine the intervention and identify its potential for scaling up or using it in other contexts. On the other hand, if a small organisation is delivering a STEM engagement activity with a small budget that has a wealth of evidence for its effectiveness, then a more light-touch evaluation will be sufficient. The type of evaluation and investment in evaluation should be proportionate to the nature of the activity, the amount, how unique it is, and whether there is already strong evidence in the area.

As far as it concerns different types of evaluation designs, in the last decade, there has been an increase in experimental studies and the use of RCTs looking into the impact of youth engagement with STEM (e.g. see Roy et al., 2021; Straw, Bamford and Styles, 2017; Husain et al., 2019). A similar phenomenon has been observed in the US. Some of the experts highlighted that an RCT for informal science education would be very difficult to implement because the STEM interventions are quite diffuse; hence it will be very difficult to design and also to decide the outcomes. Moreover, *'it is notoriously difficult to develop "rigorous" designs and methods for contexts where participants typically expect an enjoyable, non-threatening experience'* (Allen and Peterman, 2019). Using RCTs for STEM interventions at schools was considered more feasible.

> *'But you'd have to design it [the STEM intervention] in a way where you would expect to see results that would be very clear, as you know, because it would have to be over and above whatever else that school was doing.'*

RCTs have been criticised for importing a medical model into education, and that is problematic for looking at impact especially in informal science learning experiences where it is hard to control all the variables of free-choice learning determined by the young person.

> *'It's not like doing a proper RCT where you can control lots of factors. And you're only looking at one thing, which is, to see if a medicine makes people better, you know, that reduces a condition or makes a condition better to live with. You can assess that one thing. But we are not assessing one thing with our STEM engagement activities normally anyway.'*

There have been a number of issues that have been encountered in RCTs of youth engagement with STEM activities impacting the reliability and validity of their findings. Some of the issues are common to other evaluation methods, and some are unique to the RCT design. There is difficulty in getting high response rates from a large number of young people, especially when they are older and their time and attention focus on

GCSEs. There are issues of fidelity as the intervention is highly likely that it's not implemented as close to its original design. That is especially problematic when the intervention goes into an efficacy trial without having been piloted sufficiently. There are issues of contamination, e.g. when there might be parental involvement or other influences outside the school effecting the young person's attitudes and behaviours.

> *'It is important an intervention to be piloted and then to assess whether the programme was slightly premature in terms of its life cycle of where it was. In hindsight, I would say it [the intervention] wasn't really ready for RCT. There was a pilot study, but the components of this actual programme were quite different at pilot stage than they were at implementation stage. Another issue was with the delivery. When the delivery organisation started, they found that they weren't able to deliver on a scale and commissioned two other organisations. As a result, the delivery had many inconsistencies.'*

In relation to methods used to evaluate youth engagement with STEM, there is a predominance of using interviews and surveys (Fu et al., 2016; Fu, Kannan and Shavelson, 2019c), which has been confirmed by this study's review of 18 UK STEM evaluation reports. These traditional measures enable respondents to communicate their perspectives on their own and others' learning, attitudes, engagement, and behaviours (Fu et al., 2016), but they also have limitations. Fu, Kannan and Shavelson (2019c) explain that intervews and surveys rely on self-reporting, are susceptible to the reactive effects of measurement (participants may try to please the evaluator with their responses and over- or underestimate what they know or do), and they tend to interfere with the participant's learning experience. Some methods that are currently used to address the self-reporting limitations are embedded assessments or collecting data using audio- and video-recordings. Allen and Peterman (2019) point out that advancements in the use of tablets, smartphones, and go-pro cameras for evaluation purposes tend to put informal STEM evaluators in the role of detectives, searching for non-invasive ways to collect data. These developments highlight the need for revising data protection and safeguarding regulations.

There is a demand to develop new methods, especially in more informal STEM engagement environments, that will enable capturing outcomes in more direct and less obtrusive ways, allowing triangulation with multiple measures on outcomes (Fu, Kannan and Shavelson, 2019c). In addition, there is a need for new methods that go beyond measuring short-term outcomes, that can capture the complexity of learning across time and settings, aligned with the concept of the STEM learning ecosystems (Baron, 2014). Moreover, young people don't find surveys very engaging for providing feedback; hence it is important to develop more creative methods for capturing their views. Involving young people in the development and testing of these methods will enable creating more relevant and effective measures.

8. Measuring Diversity, Equity and Inclusion (DEI)

The STEM sector in the UK has, over the last decade, increased its focus on DEI issues following a wide recognition that informal and formal ***STEM*** education tend to better serve white, urban and more socioeconomically privileged young people (Godec and Archer, 2021). A number of developments indicate progress in the field, including UCL's ASPIRES project and the concept of science capital helping to address inequalities in STEM participation, the British Science Association's new strategy focusing on underrepresented and underserved audiences, and a wealth of projects aiming to increase participation in STEM from a wider and previously underrepresented group of young people. Consequently, there has been relevant progress in evaluation practice such as the creation of The Equity Compass for assessing how equitable or not specific programmes and activities are (YESTEM Project UK Team, 2020) and the provision of guidance and tools on how to measure aspects of DEI (e.g. Tomorrow's Engineers Demographic Data Question Bank).

Youth engagement with STEM efforts have been criticised in that they focus on diversifying participants by helping young people overcome issues of access and have adopted a deficit-based perspective without addressing the conditions that have created inequities. In contrast, equity-focused efforts locate the "problem" within STEM engagement institutions and aim to reconceptualize STEM engagement to reflect a broader range of cultures and practices. Evaluation has a role to play in supporting more equity-focused efforts by privileging the voices and lived experiences of non-dominant groups of young people, engaging young people in identifying desired outcomes as part of front-end evaluation, and ensuring that instruments and measures used are valid for the culture and context in which the evaluation is situated. Caution is also placed over the use of shared measures and constructs as they can become representations of normative, dominant culture practices that are then accepted as defining the intended outcomes of youth engagement with STEM. Evaluation that focuses on culture and context is often referenced as culturally responsive evaluation and requires specific awareness and sensibilities of the evaluator enhanced by shared lived experience between evaluators and young people. Therefore, there is a need for evaluators with a more diverse background.

While a lot of summative evaluations seek to understand 'what works' and compare interventions across settings, it is important to have a more expanded perspective on effectiveness by seeking to understand "what works, for whom, under what circumstances." To do this, evaluators need to continually ground their findings (including data from shared measures) in context and to look to emerging views of causality and causal methods that consider—rather than try to strip away—context, culture, and complexity. Conversations about broadening STEM participation often focus on how to attract and support students in completing degrees and seeking careers in STEM, practices associated with the STEM career "pipeline". To help advance equity and inclusive practices at the field level, evaluators should attend to the ways that context and culture constitute programs, examine how replication may or may not be appropriate for a given program, and identify what adaptations may be needed if program models are transferred across contexts (Garibay and Teasdale, 2019). One of the experts interviewed explained trends in youth engagement with STEM evaluation in relation to DEI, including how the terms are being conceptualised and looking at STEM as part of the culture.

> *'The strongest push in every area in the field is diversity, equity access and inclusion. Culturally responsive, culturally competent evaluation has become the primary order of the day, in communities engaged in the evaluation, and even in the design of evaluation from the beginning. There is a growing body of research literature on equity in STEM. There is a continuum of understanding of the terms and the terminology and what it means for practice. Some evaluators have been paying attention to this for a long time and continue to evolve their understanding of the distinctions between equity and inclusion and could tell why they are distinct and how to measure whether you are being more inclusive or equitable. A critical stance with regard to research and evaluation is necessary because it is not just measuring the STEM pipeline and what kind of activities or programmes influence youth to go into a STEM career. It's thinking more broadly about how STEM is part of culture and manifests differently in different cultures.'*

9. Evaluation culture

Issues in relation to an evaluation culture in the STEM engagement sector include sharing evaluation learnings, including failures, and being evidence-based. These issues have been persistent in the last decade and highlighted through various reports (e.g. Lloyd et al., 2012), practitioner surveys, and convenings (e.g. NCCPE, 2019, Bultitude, Verbeke and Duncan, 2015) and the expert interviews in this study.

9.1. Sharing evaluation reports and learnings

Some progress has been achieved in relation to sharing evaluation data and findings in the UK, but there have been some backward steps, too. Experts suggested that the STEM engagement sector is increasingly becoming more open to sharing evaluation findings and reports but they lack the mechanisms and infrastructure to do so. As it was aforementioned, in the US, a successful online platform that collects proactively, curates and shares evaluation reports in informal science education is informalscience.org. It is being managed by CAISE and is funded by NSF. One of the definitive decisions that helped build the repository has been that NSF requires their grantees to upload their evaluation reports on the platform as part of the formal project completion.

In the UK, the British Science Association (BSA), more than a decade ago, created Collective Memory, an online database of evaluations of public engagement with science initiatives. Some funders encouraged their grantees to share their evaluations through Collective Memory whilst others made it a requirement but without any follow up checks. Issues related to the curation of the evaluation reports, lack of clear thinking of what the platform would look like, and how it could work led to the platform being underused and eventually abandoned. As the STEM engagement field is dispersed it is often difficult to locate evaluation documents, and evaluation reports are not always publicly available (Godec and Archer, 2021). The success of informalscience.org in the US and the continued articulated need from the STEM sector for a mechanism to share their evaluations suggest the need to re-examine the development of an online hub.

> *'You can't ask people to just volunteer their information, because they won't, because we know that that doesn't work. There's got to be some infrastructure in place and some mechanism for people to do it.'*

Moreover, experts suggest that merely making evaluation reports available is not enough. It is important findings across evaluations are collated and synthesised in a format that is user friendly and accessible. Funders themselves need to be more proactive with synthesising findings across projects within their funding programmes and sharing these with the wider community.

> *'It's deeply problematic when a funder has access to all that knowledge, they've got all of that insight, but they have not presented anything back to the community. They're not saying, 'what we've learned out of all of those projects that we funded, here's what you can be doing better as a community, or here's how we've changed our funding so that you can do better practice'. And so, it frustrates me when that sort of thing happens. I have a very small grant scheme at work. And we get evaluation reports from everybody that we fund. We then synthesise those evaluation reports, we put out…guides and reports based on that. It's not difficult. It's a bit complicated, but it's not impossible'.*

Interviewees pointed out EEF's approach as successful for systematically capturing evidence and sharing it with the wider sector. EEF has been effective in terms of mobilising the knowledge base about interventions that work in schools. EEF ensures that the interventions' trials reports are accessible to practitioners and publish these using a particular format on their website. An individual trial report becomes part of a larger

suite of reports, the Teaching and Learning Toolkit, which summarises evidence on the impact of approaches tested in the past. EEF is also creating guidance complemented by tools, resources, and training to support their implementation. Moreover, NCCPE is producing guidance for public engagement with science practitioners using the 'what works' process. They choose a topic based on feedback from public engagement stakeholders, conduct desk research to identify existing published evidence on the topic, crowdsource further evidence directly from their network, host an event for experts to synthesise the learning into a draft resource and publish the final guide on the NCCPE website.

9.2. Sharing failures

One of the challenges and gaps in relation to sharing evaluation learnings related to youth engagement with STEM is the lack of sharing failures and what didn't work when a programme or an intervention got implemented. As evaluation is often conducted for accountability purposes and as a funder requirement, grantees do not feel comfortable revealing that funding has been spent on an intervention that didn't work. The issue is exacerbated when the staff involved in the delivery and evaluation of STEM engagement projects are employed through short-term contracts and when organisations rely heavily on project funding. Income uncertainty creates pressure to capture and disseminate only positive evidence of impact.

Organisations that have an open approach to sharing the outcomes of evaluation, including failures and are committed to being transparent with others so that they can learn from learnings attribute their attitude to the safety the core funding is providing them. That makes them feel confident and comfortable in sharing the evaluations, whatever they find, and also in being able to invest in them. One interviewee described how their organisation that has strong core funding is currently subsidising the programme evaluation of a long-term STEM programme and also negotiated with another funder to divert funding from expanding delivery to more young people, to evaluate and iterate. Having a trusting relationship with a funder is key to creating an evaluation culture that shares what works and what doesn't.

> *'Funders and grantees could have more open and honest relationships. But… that's about focusing on the human side of this and the culture that that's created.'*

In the Arts sector there has been a similar attitude around lacking ownership of failure, as one expert explained:

> *'There seems to be this lack of ownership around failure. And.. being open and honest about things that haven't worked, what things could be done better and why… partly that is because people are out there looking for money, and they don't want to…showcase the fact that things might not have worked, because there's a fear that they might not be able to attract further investment for that. But … we need to overcome that. And, actually, people need to be open and honest, and say, 'this didn't work and these are the ways that we're going to improve things going forward'.'*

9.3. Being evidence-based

Practices in the STEM sector are rarely conceptually informed or evidence-based (Godec and Archer, 2021). This statement aligns with the review of evaluation reports that was conducted as part of this study. The reports don't indicate whether the findings are going to be used in the future to improve an intervention. It is encouraging to see that some impact evaluations are increasingly complemented with process evaluations, but it is not clear to what extent the recommendations will be taken forward. Funders have the responsibility to become role models in using themselves evidence from their programme evaluations and share how they improve their practice accordingly.

'I think if they [funders] use the evaluation, they did that would make a step change across the sector. Because let's be honest, funders don't use it. I don't think they read them. I mean, they don't have time. I don't mean it in any disrespectful way…they asked for an evaluation because I think it's important, and they want people to self-reflect on their learning. And then they almost just kind of have a quick glance over it and go.'

EEF, in the early years of their work, faced resistance from the teachers when trying to support the sector to become more evidence-based. Teachers didn't accept that evidence could help them make better choices in the classroom and criticised RCTs as unethical. EEF tried to communicate how using evidence may be beneficial to teachers without undermining their professional expertise. As part of their efforts to change the culture towards evaluation they used appropriate language, and emphasised that teachers can look at the Toolkit's evidence, but they also need to use their professional expertise to look at it in their own settings and see if it's working for them. Making sure they respect teachers' professional expertise means now that teachers are really comfortable about using evidence. Similarly, to create a more evidence-based STEM engagement sector, it is important to understand what is holding them back from using evidence.

Two approaches have been suggested as successful for changing evaluation culture and increasing the usability of evaluation in STEM engagement and more widely. Having an external evaluator who will work with the STEM organisation, create rapport, keep them engaged throughout the process and feedback findings as they emerge or having an external evaluator who will collaborate closely with the delivery organisation, by involving its staff in collecting, analysing data and sharing the findings. Creating a close collaboration between practitioners and expert evaluators/academics is not without challenges as the experts may use language inaccessible to practitioners and create reports that are not easy to comprehend. Similar situations are observed in the Arts sector highlighting the need to facilitate practitioners' and expert academics/evaluators' interactions in order to develop an understanding of how each works and communicates.

'Academics are hugely intelligent people who …use complex language to talk about… various evaluations, initiatives, etc. If you cannot translate that in an easy-to-understand way, … for the cultural sector to understand or you cannot engage them in a way that they are interested, then you've lost…there's just no interaction happening whatsoever. … That's probably one of the biggest issues;…you've got people with expertise, coming into organisations, undertaking evaluations, and then leaving the cultural organisation on its own. And, you know, alright, fine, and evaluation tick, there is a report, that's an evaluation now, is there any learning? Is there any understanding that's been left with any of the people who are working within that organisation? Probably not… One of the biggest … issues going forward is how to get academics to understand how… the cultural sector is working. 'And how do you get academics not to pursue their own particular interests, but to actually tackle real world cultural problems within the evaluation side of things that …will actually make a big difference to the sector and the way in which evaluation is undertaken?"

10. Training, support networks and partnerships

There is no formal training dedicated to evaluating youth engagement with STEM. Training may be offered e.g. for museum professionals as an elective course as part of an MSc course or sporadically training is being offered by large STEM engagement providers such as the Science Museum Group or networking organisations such as NCCPE and the Visitor Studies Group. In addition, there are not many professional development opportunities for professional evaluators, focusing on evaluation of STEM engagement. Some more generic evaluation training is being offered by professional bodies such as the UK Evaluation Society and the Social Research Association.

It is being suggested that advancing evaluation of youth engagement with STEM requires building the capacity of organisations to conduct in house evaluations or alternatively to build their knowledge on how commission evaluation externally. Moreover, funders could support improving evaluation literacy by offering relevant training courses to their grantees. Any evaluation training offered should also take into consideration the staff turnover in STEM organisations and should be delivered regularly rather than considered as a one-off offer.

There seems to be a disparity in the evaluation capacity in the field. For example, big museums with a lot of resources have more capacity, skills and resources for evaluation. Smaller organisations and individuals are lagging behind in terms of evaluation skills and understanding. Accordingly, evaluation training needs to be differentiated to be able to cater for different levels of need and capacity. The need for specialised training in evaluating STEM engagement is indicated by the popularity of registrations for evaluation training in the sector, e.g. NCCPE training.

There is a demand to equip the people who deliver youth engagement with STEM to do their own effective evaluations. Realistically it may not be possible to have staff dedicated to evaluation. Hence, it is important to equip the staff who deliver STEM activities on how to do this.

Establishing collaborations between practitioners and academics/evaluators, e.g. as part of a project evaluation, can also involve capacity building by offering training as part of the project. In addition, mentoring type of support by experts on evaluation throughout a project has been mentioned as an approach to increase the evaluation knowledge and skills of delivery staff. An example of a recently set up dedicated mentoring scheme is the Visitor Studies Group Mentorship Programme. The programme offers support to emerging audience researchers or anyone who has had audience research responsibilities added to their role. Long-standing VSG members become mentors, and the programme is tailored to the needs of the mentee, from the frequency of meetings to the objectives set up to be achieved.

11. Challenges and opportunities for evaluation due to Covid 19

During the Covid 19 pandemic, the closures of schools and informal science learning settings had an impact on the provision and evaluation of youth engagement with STEM. STEM interventions that were evaluated through RCTs had to stop or be redesigned significantly as it wasn't possible to deliver them as they were originally designed. Similarly, longitudinal studies came to a halt. Also, importantly provision suddenly moved online and consequently, there was an urgent need to capture impact using new appropriate methods that were developed for virtual STEM activities. Methods that require face to face presence, such as observations, had to be adapted whilst using surveys online with young people often didn't work well as young people don't want to complete digital surveys. On the other hand, running online focus groups with young people as part of the evaluation became easier, as barriers such as the need for transport and more time for travel to the venue were removed. Conducting evaluation online, e.g. using zoom, provides flexibility in different ways. People with disabilities encounter fewer access issues when they participate online and young people may opt to have their cameras off and use the chat function to write their feedback if they prefer it, instead of talking. Methodologies such as digital ethnography gained popularity.

When the provision moved to digital it was much easier to capture how many young people signed up. An expert from the youth sector highlighted that there were opportunities to do evaluation differently/better but they weren't taken. For example, when young people were signing up on zoom sessions there was a missed opportunity to track numbers of young people but also socio-demographic data regarding their background.

Another challenge was related to evaluation capacity. Not many external evaluators were experts in evaluating digital provision before the pandemic and many STEM engagement organisations didn't have the expertise or the staff to evaluate online engagement in house. Additional ethical issues emerged for evaluating online engagement and demand for new ethical considerations guidance.

On a different note, Covid 19 provided an opportunity to conduct natural experiments, for example to compare face to face delivery of a programme pro-pandemic as opposed to delivering it online. Evaluations have already started to emerge in the STEM field looking into such comparisons and whether a hybrid version of a programme is more effective moving forward (e.g. see evaluation of Nuffield Future Researchers (CFE Research, 2021)).

Experts also suggested that Covid 19 strengthened the case for evaluation and brought more attention to DEI issues. The sensitivity to DEI may have been caused because during the pandemic, with the murder of George Floyd and the Black Lives Matter movement people were made to stop and think about systemic issues that are excluding people.

> *'I think that it [the pandemic] has made people care more about return on investment. And it's made people care so much more about young people from disadvantaged groups. And actually, they are starting to think, 'ha maybe it isn't working as well for them. I need some evidence.' So maybe it has helped make the case for evaluation stronger.'*

12. Recommendations

In the UK, the youth engagement with STEM sector is varied, affected by many factors and different funders. As this report has shown, the sector has demonstrated progress in evaluating its practice in the last decade, but persistent challenges and gaps still need to be addressed. In addition, different groups in the sector are working separately towards improving evaluation. Bringing change in the STEM sector as a whole requires an organization to take the lead and unify the dispersed efforts. Based on the desk-based research and interviews by experts from the STEM, Arts and Youth fields, the following recommendations are proposed for how funders and policy-makers can support the STEM sector to improve its evaluation practice.

12.1. Convening and collaboration between funders and policy-makers

Across the UK, the government and other policymakers and funders have a strong interest in improving and investing in the STEM sector. A convening funder or policy maker with credibility, networks and power can bring these together to look at what's driving and motivating their investments, how they are evaluating these, what evidence they need, who their internal stakeholders are and create a coalition that will support the STEM sector to improve its evaluation practice in relation to youth engagement. In addition, funders and policy-makers are influential in the STEM practice sector and grantees are motivated to produce a good quality evaluation of their funded projects.

A convening could involve creating a shared funding programme to improve youth engagement with STEM evaluation, following a similar model to the Science Learning+ collaboration between Wellcome, the Economic and Social Research Council in the UK and the NSF in the US.

Coordination of STEM funders and policy makers could also involve bringing them together, e.g. UKRI, The Nuffield Foundation, The National Lottery Community Fund, and others, to align in terms of requirements for their grantees such as outcomes and project evaluation. This will lift the burden for grantees that often have to develop a new approach and report for a different set of outcomes for different funders, and it will enable more effective use of the evaluation resources. A similar approach has been followed by the Institute for Voluntary Action Research (IVAR) that convened charities and funders to agree and commit to common principles and approaches to managing grants and relationships with grantees. The result has been the development of the flexible funders' commitments.

12.2. Encourage debate, discussion and bring consensus in the sector

Experts interviewed for this study noted a need in the STEM sector for a wider discussion and consensus in relation to the purpose of engaging young people with STEM before even discussing how to improve the evaluation of the activity. The stakeholders involved in and influencing youth engagement with STEM vary, i.e. from young people, government, local authorities, and funders to museums, schools, STEM employers and universities, just to name a few. As the sector is interested in demonstrating their impact collectively, they first need to articulate more explicitly what their motives are, why they are engaging young people and what they are trying to achieve.

Creating a shared common outcomes framework followed by a ToC in the sector has been one of the recommendations for the next step in this process. Funders and policy-makers have the resources and level of influence to bring together the sector to discuss and potentially build consensus on what is/are the purpose(s) of engaging young people with STEM, build an outcomes framework and facilitate a process for creating an overarching ToC. Professionals who deliver STEM activities and evaluators can then receive training on how to use the framework, and a ToC and funding will be needed to enable collaborations between evaluators/researchers and delivery staff to evaluate programmes and test the framework. These projects could be then used as examples of how the framework and ToC can be applied in practice and convince others in the sector to use it. Different STEM networks, e.g. NCCPE, VSG, EngineeringUK, can help their peers, offer training, influence and increase the take-up of this outcomes framework and ToC. Each group could also be supported to create their own ToC. Working with stakeholders across the sector and their networks is crucial. Stakeholders need to be involved in shaping a common outcomes framework and a ToC so that they will feel ownership and commitment to using them. Moreover, a ToC for STEM engagement could show the common outcomes and common pathways to increasing STEM engagement and the different stages of the pathways. That might give people a lot of support to develop their own programmes, specific ToC, and select outcomes.

As some experts cautioned, different groups in the sector are already using this approach and have created ToC for projects, initiatives, or for a particular subgroup of STEM. Hence if an overarching ToC is to be developed, it needs to consider and build on current practice and experience. This seemingly daunting and difficult approach could enable the sector to *'move on from little working groups, talking about how to improve and making little bits of improvement here and there to a system improvement'*.

12.3. Support the development and use of ToC at project and initiative levels

Funders and policy-makers could also support the STEM sector to use ToC to improve their practice. Through funding and the development of resources and case studies, the STEM sector can be supported to create ToC for current and future engagement projects and test the ToC. By conducting an implementation and process evaluation, it will be possible to find out if changes occur or not. If change is not visible, then changes in the programme or the ToC may be needed. If the evaluation has shown that the ToC is working, then further testing of the intervention, e.g. through an RCT, might show changes and progression towards the outcomes in the longer term. An example from the US of a project funded by the NSF to develop a ToC for a STEM engagement activity and determine the feasibility of conducting a national scale study is Evaluating STEM Scouts: The Design of a Comprehensive Evaluation Plan and Feasibility Study

12.4. Support the robust data collection to track change and impact and increase access to shared measures

Stakeholders in the STEM sector have expressed the importance of being able to tell as a sector their collective story of impact over time. In order to enable this, funders and policy-makers could consider investing in an online hub co-created by the STEM sector. The online hub will include tools for measuring impact based on the overarching agreed ToC, training for organizations and individuals to use these and an online data collection portal for submitting the data. As a return, the platform will provide analysis of the data and enable comparison and collation across the sector. ACE has already followed this approach by investing in the Impact and Insights toolkit, delivered by Counting what Counts. The toolkit provides the arts and cultural organizations with a shared approach to evaluating their work. The toolkit uses sets of metrics to explore what audiences, participants and peers think about a performance, exhibition or project. The resulting insights can be used to inform future practice whilst helping build a greater understanding of the cultural sector.

The online hub could also include a section that will provide access to a wider set of shared measures similar to the ATIS repository of measures in the US. Even though there are a number of projects producing shared measures for the youth engagement with STEM field, there are barriers to accessing some measures. Some evaluators consider the instruments proprietary and are reluctant to share them with others. Others publish instruments in peer-reviewed journals that are often inaccessible to evaluators. Additionally, there are some projects that charge a fee to access an instrument. However, fees may be needed to develop structures related to the use of a measure, such as a shared online platform, database, and reporting features (Grack Nelson et al., 2019). Overall, the future of shared measures will require changing mindsets around access to individual tools and the creation of the online hub will enable easy and user-friendly access to a range of measures.

An effort to collect data across the sector consistently requires an agreement to shared measures. Funders and policy-makers can join forces and convene the STEM sector in the UK to discuss and try to come to a consensus of what shared measures could look like that could be used across the sector. In the US in 2019, the NSF funded the From Common Measures to "Measures in Common" convening. The conference brought together 72 practitioners, researchers, evaluators, and other stakeholders to explore the current state of evaluation and measurement tools in afterschool STEM programmes and address the need to monitor the quality and outcomes of a wide range of programmes.

12.5. Support the more open exploration of outcomes of youth engagement with STEM activities

Critics to the efforts to create overarching outcomes frameworks and ToC across the sector argue that these approaches create a narrow perspective of what STEM engagement achieve. Instead, funders and policy-makers could invest in creating rich, well-designed activities and using an ethnographic methodology to explore all the possible impacts. Having a more open approach to exploring the outcomes that occur will enable the identification of a wider variety of the most common outcomes rather than trying to get the lowest common denominator.

12.6. Invest in long-term evaluations of key youth engagement STEM programmes

Funders and policy-makers could play a major role in supporting a longitudinal piece of evaluation and coordinating the sector to be more coherent in the measures that they are using and the purposes for using these. Significant investment is required for delivery organizations to gather consistent data over time. There are various reasons for investing in such an effort, i.e. creating a longitudinal perspective on how to effectively engage with young people, evidencing the engagement's impact and convincing the policymakers that these delivery mechanisms work. It should also be acknowledged that bringing change, e.g., increasing diversity and inclusion in youth engagement with STEM, takes time, and it is important to set up long-term evaluation that will track this progress.

Investing in longitudinal studies can be of different forms, including:

- Setting up a longitudinal/cohort study to track and monitor youth engagement with STEM over time and identify issues with engagement, a similar approach to the Wellcome's Science Education Tracker.
- Setting up longitudinal studies of programmes that have been delivered over the years but haven't been evaluated sufficiently to evidence their impact. Evaluation of programmes delivered at scale can result in identifying issues with the delivery and enable improving practice.

12.7. Influence the evaluation culture

A key issue in improving evaluation practice is whether and how evaluation is used and shared. There are two areas funders and policy-makers can focus on in order to change the culture towards evaluation in the STEM youth engagement sector.

12.7.1. Enabling and encouraging sharing evaluations more openly

There is a clear need for more sharing of evaluation reports so that others can learn from the findings of project evaluations. In the US, the NSF has supported, for almost two decades, the informalscience.org repository of evaluation reports from the informal science learning sector. A critical factor for the repository's success has been NSF's strong stance, requiring grantees to upload their evaluation reports on the platform as part of their project completion process. Funders and policy-makers may consider funding a similar platform that will focus on youth engagement with STEM activities and projects and require its grantees to upload their evaluation reports on the repository. The BSA Collective Memory platform is now closed due to issues related to the curation of the reports and the sector not being evidence-based. That means close consideration needs to be made if a repository will be developed so that it is co-created by the organizations who will use it, it is user-friendly, and incentives are given to the sector for using it.

It is recommended that the repository also offers access to relevant STEM engagement academic journals, which often include publishing papers and evidence from the evaluation of STEM projects and are currently closed access behind paywalls. Informalscience.org offers its free members access to a large number of journals thanks to investment from NSF, which supports offering the field access to research.

Funders and policy-makers, apart from offering a repository to share evaluation, can also play a more active role in sharing evaluation findings, e.g. through case studies using effective communication channels, e.g. social media, practitioner and academic journals, and conferences. By disseminating effectively evidence of successful youth engagement with STEM, funders and policy-makers can also encourage the sector to use the evidence and learnings in future initiatives.

12.7.2 Using evaluation evidence to improve practice rather than accountability

Often delivery organizations evaluate an activity for accountability reasons and to satisfy funders' requirements. Funders and policy-makers can join forces and encourage the STEM engagement sector to conduct evaluation as part of reflective practice. This could be done in the way funding calls frame evaluation requirements by having clear expectations that reflective, evaluative practice needs to be built from the start of the project. Often it may be the grantees' perception rather the funders requiring evaluation for accountability purposes. This is why the funders need to take a stronger stance and encourage and resource project evaluations adequately for learning purposes. This will also require funders to establish more open and honest relationships with their grantees so that they will feel safe to report and publish failures and what didn't work during the project implementation. Creating clear evaluation guidance for grantees that explains what evaluation is for and the funder's expectations is important for moving the sector beyond the accountability tick the box mentality around evaluation.

Funders and policy-makers can create awards to celebrate good practice in evaluation and showcase through case studies examples of projects that share with equal emphasis what worked and what didn't.

> *'There might be a certain level of resistance out there about people understanding the impacts of their own work and accepting failure. Some people, because of public money or whatever, don't want to admit that some things have actually failed. We've got to get away from all that; we've got to improve the culture; that it's okay to fail. And it's okay to sort of better understand different initiatives what difference are actually making, and these are the reasons why it's important to do that.'*

Moreover, as part of encouraging a more evidence-based sector, funders and policy-makers may require project applications to reference previous evidence that they have based the development of their youth engagement activities with STEM. It is also recommended that funders and policy-makers advocate for a more formative and front-end evaluation model. STEM engagement practitioners need to be supported and incentivized to use more front-end and formative evaluation rather than focus solely on the summative results at the end. Funders and policy-makers can also carefully design their funding programmes that invest in youth engagement with STEM so that evaluation has increased importance and is supported appropriately. Support can be through the allocation of sufficient budget, ensuring evaluation is factored from the beginning of the project, and providing tailored evaluation guidance to grantees through an in-house evaluation expert to advise, e.g. on proportionality of evaluation, relevant methods etc.

12.8. Create the bridge between academics and STEM practitioners

Funders and policy-makers can act as a bridge to get academics to understand how the STEM engagement sector is working and encourage academics not to pursue their own particular interests but to tackle real-world issues relating to engaging young people with STEM. In the Arts sector, it was highlighted that often experts come into organizations and undertake evaluations, but then they leave the organizations on their own, with a report and not a clear understanding in the organization of what the learning was.

In addition, creating shared measures requires work beyond silos, and it demands various types of expertise that may not reside with a single evaluator. For this reason, evaluators, academics, and STEM delivery practitioners need to work together to define outcomes, develop items and protocols, and test measures (Noam and Shah, 2013). It should be acknowledged that creating partnerships between academics and STEM practitioners is a process that takes time and resources.

12.9. Invest in piloting innovative methods to measure engagement

There is a well-articulated need in the sector to improve the tools and methods for measuring youth engagement. Funders and policy-makers could launch dedicated funding programmes for practitioner-researcher/evaluator collaboration to create new ways to capture the impact of youth engagement with STEM. This should consider, for example, the need to move beyond self-reporting methods to measuring engagement in real-time and the need to evaluate effectively online engagement. In particular due to Covid 19, there has been an increased demand to use more online methods to evaluate youth engagement, such as digital ethnography or behavioural science techniques. Investment in piloting new evaluation methods could also require the active involvement of young people throughout the process, from deciding what to evaluate to prototyping and testing a new method so that it will be more relevant, as unobtrusive as possible and robust.

New assessment and evaluation technologies require careful procedures around informed consent, privacy, and data security (Fu, Kannan and Shavelson, 2019c). The use of facial and gestural recognition, location tracking, social media, surveillance cameras, big data, and more, raises concerns about how and from whom data should be collected and analysed (Fu, Kannan and Shavelson, 2019b). Funders and policy-makers have the responsibility and resources to ensure ethical guidelines are updated to include considerations when applying new evaluation and assessment technologies.

12.10. Support improving measuring DEI and diversifying the evaluation sector

There is growing interest and attention in STEM, similar to other sectors, in terms of DEI. Although many discussions have focused on diversifying access to STEM careers, increasingly, the focus has become wider to encompass decolonizing STEM. The rise of the concept of science capital and the equity compass have provided a lens and tool for professionals and their organizations to use and reflect on their practice. There is a need to support the sector to clarify what the terms DEI mean in their practice and to develop tools to effectively measure DEI in youth engagement with STEM. For this to happen, it is important to convene discussions in the sector to map the difficulties they find in measuring DEI, identify examples of best practice, and, if possible, agree on some shared measures to enable tracking in the long-term on how they progress in becoming a more DEI sector. Equally important in improving the evaluation practice is increasing the diversity of the evaluation professionals. In the US, one such initiative is the Graduate Education Diversity Internship (GEDI) Programme set up by the AEA. The program provides internship and training opportunities to support graduate students from groups traditionally under-represented in the field of evaluation. CAISE has been connecting with the programme, which gives an opportunity, especially for smaller STEM organizations to work alongside a graduate evaluator who is also receiving mentoring and help from GEDI. Funders and policy-makers can play a leading role in funding and convening STEM stakeholders to discuss and co-create effective methods to measuring DEI and to invest in initiatives dedicated to diversifying the pool of evaluators.

12.11. Support communities of practice and networks on improving evaluation practice

Funders and policy-makers can support practitioner networks to bring together professionals to share evaluation findings, best practices in evaluation and failures. These can be existing networks dedicated to the evaluation of engagement with STEM, or if needed, new networks could be formed with a focus on improving evaluation. For example, the ACE's Impact and Insight Toolkit supports a peer network focused on evaluation. These networks don't need to be broad; experts highlighted that smaller trusted networks where professionals are committed to their practice, feel safe and can begin to have open conversations about evaluation can be more effective. Technology and the means to meet up virtually make it easier for such networks to meet without requiring much time and financial resources.

12.12. Invest in the development of evaluation resources

There is a wealth of evaluation resources for assessing youth engagement with STEM, but they are dispersed in different hubs and sources. In addition, there are gaps in the available resources. Funders and policy-makers can invest in a sector-wide convening to identify what types of resources are needed to improve evaluation practice, identify and collate existing resources that can meet the needs and commission the development of new resources where there are gaps. A similar approach has been followed in the US. In 2013, with funding from NSF, CAISE held a convening (Ellenbogen, 2014) to facilitate discussion about the resources needed to improve evaluation quality in ISE. An example of a very popular resource developed in the US has been the Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects, an initiative by CAISE and the Visitor Studies Association (Bonney et al., 2011). It was developed by a collaboration between evaluators and practitioners who got together and aimed at the designers of interventions and programmes on how to work better with evaluation consultants. A similar resource could be created in the UK to fulfil the need to improve how funders and STEM organizations commission evaluations and work with consultants. What is crucial in developing new resources is to create them from a user perspective. It is important to make

the resources available through a centralized hub, e.g. the same that may include a repository of evaluation reports. In the US, informalscience.org, apart from being a repository of evaluation reports, it is also a platform for identifying informal STEM education projects funded by NSF and other funders. The website also includes blogs, project highlights, a newsletter, a calendar of events, all related to learning in informal settings, and resources for proposal development, research and evaluation, which are very popular and are regularly visited. Existing and new resources need to be made accessible but also disseminated more widely, e.g. through webinars and training to increase the possibility to be used.

12.13. Invest in evaluation training

There is a need for the provision of training on evaluation of youth engagement with STEM. Funders and policy-makers can provide evaluation training as part of the capacity building of their grantees or training could be delivered by other credible organisations and networks and offered more widely to the STEM sector. Training needs to cover a variety of needs and levels of skills and knowledge on evaluation. Training can be in the form of professional development for experienced evaluators or it can introduce basics around evaluation to inform staff in STEM engagement organisations that may commission evaluation externally. Due to staff turnover in STEM organisations it is important to repeat regularly the sessions to target new members of staff. Organisations such as VSG, Museum Association and NCCPE are well-positioned to develop and offer training, provided they are given appropriate funding.

12.14. Equip the sector with resources/tools to self-evaluate

Funders and policy-makers can support the creation and dissemination of tools that will help professionals who engage young people with STEM to self-evaluate their practice. These can include resources on how to develop a ToC, tools for testing a ToC, and tools for conducting different types of evaluation, including RCTs. For instance, EEF produced, in 2013, the very popular DIY Evaluation guide for schools (Coe et al., 2013) which supports teachers to conduct small-scale evaluations in schools. Helping practitioners who deliver youth engagement with STEM to conduct their own effective evaluations can contribute to changing attitudes towards evaluation, making evaluation less intimidating as a process, and can also encourage more innovation. Professionals will be better equipped to try a new method of engagement, test it using self-evaluation tools and refine it. Bequette et al. (2019) suggest paying attention to adapting carefully robust tools designed for trained evaluators so that STEM practitioners can understand and apply them in their projects.

12.15. Commission studies

Experts have suggested commissioning the following studies to inform further and support the improvement of the evaluation of youth engagement with STEM:

- Systematic review or synthesis of evidence on what works in youth engagement with STEM and identification of gaps.
- Practice review of the youth engagement with STEM sector in and out of the school to capture what type of activities/initiatives organizations, schools, individuals run to engage young people with STEM and how they evaluate these (this could be in the form of a State of Nation survey or a study similar to the 2012 Wellcome Review of Informal Science Learning).
- Study to understand and learn from how other countries evaluate youth engagement with STEM.
- Study to identify the different approaches and examples of best practice in measuring DEI in youth engagement with STEM field.
- Study to understand better how to change the culture towards evaluation in the STEM sector. How is it possible for the STEM sector to use and act on evaluation findings? In what ways can evaluation be used strategically to enhance policy and practice?

12.16. Consider two approaches to investing in evaluating youth engagement with STEM

Bringing systemic improvement of evaluating youth engagement with STEM requires long-term commitment and significant convening and joining forces across the many stakeholders in the sector. There are two different approaches funders and policy-makers could consider adopting to coordinate the sector effectively. One approach would be a collaboration of funders and/or policy-makers, to create a large-scale joined funding programme that will call for proposals related to what the sector has identified as needed to improve the evaluation of youth engagement with STEM. The programme could include funding some of the suggested activities identified in this scoping study. The NSF's Advancing Informal STEM Learning (AISL) program in the US offers wide programme solicitations (calls for proposals) to ensure that it is field-driven and allow the informal STEM learning sector to submit project proposals based on their needs and not what the funder thinks they need. AISL published in 2021 its most recent programme solicitation that presents AISL's broader aims and specifies that it will fund six types of projects: (1) Pilots and Feasibility Studies, (2) Research in Service to Practice, (3) Innovations in Development, (4) Broad Implementation, (5) Literature Reviews, Syntheses, or Meta-Analyses, and (6) Conferences.

A second overarching approach funders and policy-makers could consider adopting is co-funding the set-up of a centre on improving youth engagement with STEM evaluation. Setting up a dedicated centre will make it easier for STEM stakeholders to access resources and to support them efficiently. It is also important to set up an independent entity to advocate for improving evaluation rather than the funder directly trying to influence the sector. STEM organisations and professionals are more likely to be influenced to improve their practice by their peers rather than because of a funder's requirements.

Indicative activities the centre could undertake include:

- Co-create with the STEM sector a Youth engagement outcomes framework and an overarching ToC and train practitioners in the use of it
- Collate measures of evaluation and stimulate discussion for measuring collectively and consistently

impact of youth engagement with STEM

- Maintain a repository for sharing evaluation reports
- Synthesize existing evidence on youth engagement with STEM and make it accessible to the STEM sector e.g. through guidance
- Collate existing and create new evaluation guidance and resources
- Support the STEM sector to develop skills in evaluation and reflective practice.
- Create a network of practitioners and researchers for sharing evaluation findings, best practices, and discussing how to improve evaluation practice
- Stimulate debate and discussion on evaluation
- Offer funding for evaluating youth engagement with STEM (e.g. long-term evaluations of youth engagement with STEM programmes, piloting innovative methods of evaluating youth engagement)

Last, there are different models funders and policy-makers could explore adopting for this centre (see case studies). The What Works Centre model (such as EEF) has been suggested because of its holistic function of synthesizing current evidence of what works and identifying gaps, commissioning interventions to generate new knowledge in the area of the gaps and disseminating and mobilizing knowledge. However, some critics have pointed out that the What Works Centres' evidence reviews adopt a narrow scope of what counts as best practice (privileging quantitative studies), they are limited to narrow research questions, and prioritize evidence from RCTs. A second type of model to consider is a resource centre, similar to the US-based CAISE, which is funded by NSF and manages the informalscience.org website. A third model is based on the recently established Centre for Cultural Value co-funded by Arts and Humanities Research Council (AHRC), Paul Hamlyn Foundation and Arts Council England, which was designed and set up based on reviewing different types of centres and networks through a feasibility study.

Case study | EEF – a What Works Centre

EEF, was founded in 2011 originally with a £125M grant from the Department for Education. It is an independent charity dedicated to breaking the link between family income and educational achievement. The EEF supports teachers and senior leaders in finding ways to close the attainment gap by summarising the best available evidence and generating new evidence showing what works. Its main activities can be described as:

- Publishing guidance reports summarising clear recommendations for teachers based on the best evidence available.
- Managing the Teaching and Learning Toolkit which summarises the findings of more than 13,000 trials around the world
- Funding interventions designed to generate new evidence of what works, evaluated through RCTs.
- Coordinating the Research Schools Network which is a collaboration between the EEF and the Institute for Effective Education. The network of schools has been created to support the use of evidence to improve teaching practice.

Case study | CAISE – a resource centre

CAISE was set up in 2007 and has been continuously supported by NSF, to provide infrastructure, resources, and connectivity for informal STEM educators, researchers, evaluators, and other interested stakeholders in the US. CAISE has a website www.InformalScience.org, with over 8,000 resources, including project descriptions, research literature, evaluation reports, and other documents related to quality, evidence-based informal STEM learning work. CAISE also organises task forces, inquiry groups, convenings, and principal investigator meetings designed to facilitate discussion and identify needs and opportunities for those who design for, research, or evaluate informal STEM learning experiences and settings. CAISE is providing forums and spaces for these communities to come together to address the common challenges of sustaining connections between practice and research, building understanding of and capacity for evaluation and measurement, and broadening the participation of underrepresented groups in STEM. The CAISE community currently numbers over 4,000 members from 50 countries. The resource centre has had enormous influence on evaluation priorities and capacity-building connections across individuals, programs, and organizations (Fu et al., 2019a).

Case study | Centre for Cultural Value

In collaboration with the Paul Hamlyn Foundation and AHRC, ACE set up the Centre for Cultural Value. The national research centre has five-year funding, and it is based at the University of Leeds. Other core partners are The Audience Agency, The University of Liverpool, The University of Sheffield and Queen Margaret University, Edinburgh. The centre was set up based on the AHRC Cultural Value Project (Crossick and Kaszynska, 2016) that looked into why the arts and culture matter and how they capture their impact. A follow up scoping study (Kaszynska, 2017). involved consultation with 200 stakeholders, and identified that one of the biggest and most pressing challenges in understanding cultural value is creating communities of interest and practice across these sectors.

The centre works alongside cultural practitioners and organizations, academics, funders and policymakers to:

- Summarise existing evidence to make relevant research more accessible
- Support the cultural sector to develop skills in research, evaluation and reflective practice
- Convene discussions around questions of cultural value
- Shape policy developments
- Offer funding for research partnerships through the Collaborate fund

For ACE, it was important that they supported an independent organization, i.e. the Centre to connect to individual artists and organizations, to interpret evaluation studies and present these in a way that is easy to understand and relevant to the activity in the arts sector, that would encourage art practitioners to look deeper into evaluation and why it is necessary.

References

Allen, S., Campbell, P. B., Dierking, L. D., Flagg, B. N., Friedman, A. J., Garibay, C., Korn, R., Silverstein, G. and Ucko, D. A. 2008. Framework for Evaluating Impacts of Informal Science Education Projects. Report from a National Science Foundation Workshop. The National Science Foundation, Division of Research on Learning in Formal and Informal Settings.

Allen, S. and Peterman, K. 2019. Evaluating informal STEM education: Issues and challenges in context. In A. C. Fu, A. Kannan, and R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 17–33.

Archer, L., DeWitt, J., and Dillon, J. 2014. 'It didn't really change my opinion': Exploring what works, what doesn't and why in a school science, technology, engineering and mathematics careers intervention. Research in Science & Technological Education, 32(1), 35-55.

Archer, L., Dawson, E., DeWitt, J., Seakins, A. and Wong, B. (2015) "Science capital": a conceptual, methodological, and empirical argument for extending bourdieusian notions of capital beyond the arts. Journal of Research in Science Teaching, 52 (7), 922-948. Available at https://centaur.reading.ac.uk/69983/ [Accessed 20 March 2022]

Archer, L., Moote, J., MacLeod, E., Francis, B., & DeWitt, J. 2020. ASPIRES 2: Young people's science and career aspirations, age 10-19. London: UCL Institute of Education.

Archer, M. O. 2016. So you are looking to run a research schools project? Practical tips from the Evaluation of a pilot programme. London: Queen Mary University of London.

Archer, M. O. and DeWitt, J. 2021. 'Thanks for helping me find my enthusiasm for physics': the lasting impacts 'research in schools' projects can have on students, teachers, and schools. Geosci. Commun. 4, 169–88.

Archer, M., DeWitt, J., Davenport, C., Keenan, O., Coghill, L., Christodoulou, A., Durbin, S., Campbell, H. and Hou, L. 2021a. 'Going beyond the one-off: How can STEM engagement programmes with young people have real lasting impact?'. Research for All, 5 (1), 67–85.

Archer, M. O., DeWitt, J., Thorley, C., and Keenan, O. 2021b: Evaluating participants' experience of extended interaction with cutting-edge physics research through the PRiSE "research in schools" programme, Geosci. Commun., 4, 147–168.

Baron, B. 2014. Formative assessment for STEM learning ecosystems: Biographical approaches as a resource for research and practice. Report prepared for the National Research Council Committee on Out-of-School Time STEM. Washington, DC, USA.

Bell, J. 2020. From Common Measures to 'Measures in Common' Convening. Available at: https://www.informalscience.org/news-views/common-measures-measures-common-convening [Accessed 10 March 2022].

Bequette, M., Cardiel, C. L. B., Cohn, S., Kollmann, E. K., and Lawrenz, F. 2019. Evaluation capacity building for informal STEM education: Working for success across the field. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 107–123.

Bonney, R., Hellenga, R., Luke, J., Marcussen, M., Palmquist, S., Phillips, T., Russell, L., Trail, S., and Yalowitz, S. 2011. The Principal Investigator's Guide: Managing Evaluation in Informal STEM Education Projects. Center for Advancement of Informal Science Education. Available at: https://www.informalscience.org/sites/default/files/caisevsapi_guide.pdf [Accessed 10 March 2022].

British Science Association (BSA). 2018. British Science Festival Evaluation report. BSA.

Bultitude, K., Verbeke, M. and Duncan, S. 2015. Evaluation support and practice within the National Forum for Public Engagement in STEM: Scoping study findings – Executive Summary.

Cilauro, F. and Paull, G. 2019. Evaluation of Nuffield Research Placements: Interim report. London: Nuffield Foundation.

CFE Research. 2021. An independent evaluation of Nuffield Future Researchers. London: Nuffield Foundation.

Coe, R., Kime, S., Nevill, C., and Coleman, R. 2013. The DIY evaluation guide. London: The Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation_Guide/EEF_Evaluation_DIY_Evaluation_Guide.pdf [Accessed 5 March 2022].

Connolly, M. R., and Seymour, E. 2015. Why theories of change matter (No. WCER Working Paper No. 2015-2). Accessed at: https://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2015_02.pdf [Accessed 20 February 2022].

Crossick, G., and Kaszynska, P. 2016. Understanding the value of arts & culture. The AHRC cultural value project. Swindon: Arts and Humanities Research Council.

Davenport, C., Dele-Ajayi, O., Emembolu, I., Morton, R., Padwick, A., Portas, A., Sanderson, J., Shimwell, J., Stonehouse, J., Strachan, R., Wake, L., Wells, G., and Woodward, J. 2021: A Theory of Change for Improving Children's Perceptions, Aspirations and Uptake of STEM Careers, Res. Sci. Educ., 51, 997–1011.

Denny, R. 2021. Intensive STEM Summer Camps Interim Impact Report. The Charity Spark. Available at: https://www.stem.org.uk/sites/default/files/pages/downloads/STEM%20Summer%20Camps%202021%20-%20Interim%20Impact%20Report%20December%202021%20-%20single%20page%20view_0.pdf [Accessed 16 March 2022].

DeWitt, J. 2019. I'm a Scientist: Supporting Science Capital. Report. Available at: https://about.imascientist.org.uk/files/2019/11/IAS-Science-Capital-Main-Report-Sep-2019.pdf [Accessed 18 March 2022].

Ellenbogen, K. 2014. Summary of the CAISE Convening on Building Capacity for Evaluation in Informal Science, Technology, Engineering and Math (STEM) Education. Washington, DC: Center for Advancement of Informal Science Education (CAISE). Available at: http://informalscience.org/research/ic-000000-010-034/ECB_Convening_Summary [Accessed 25 March 2022].

Emembolu, I., Padwick, A., Shimwell, J., Sanderson, J., Davenport, C., and Strachan, R. 2020. Using action research to design and evaluate sustained and inclusive engagement to improve children's knowledge and perception of STEM careers, International Journal of Science Education, 42(5), 764-782.

EngineeringUK 2021. Impact Framework for Engineering Outreach Webinar – Tomorrow's Engineers. [online] Available at: https://www.tomorrowsengineers.org.uk/improving-practice/resources/engineeringuk-impact-framework-for-engineering-outreach-webinar/ [Accessed 18 March. 2022].

Fu, A.C., Kannan, A. Shavelson, R.J., Peterson, L. and Kurpius, A. 2016. Room for Rigor: Designs and Methods in Informal Science Education Evaluation, Visitor Studies, 19(1), 12-38.

Fu, A. C., Kannan, A., and Shavelson, R. J. 2019a. Editors' notes. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 7–15.

Fu, A. C., Kannan, A., and Shavelson, R. J. 2019b. Synthesis of issues and future directions. In A. C. Fu, A. Kannan, and R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 125–131.

Fu, A. C., Kannan, A., and Shavelson, R. J. 2019c. Direct and unobtrusive measures of informal STEM education outcomes. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 35–57.

Gammon, B. 2012. Hands-On DNA: Exploring Evolution Evaluation Summary Report. London: Ben Gammon Consulting.

Garibay, C., and Teasdale, R. M. 2019. Equity and evaluation in informal STEM education. In A. C. Fu, A. Kannan, and R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 87–106.

Godec, S. and Archer, L. 2021. Informing UKRI's STEM Inspiration and Youth Engagement work: A Think Piece. Unpublished.

Grack Nelson, A., Goeke, M., Auster, R., Peterman, K., and Lussenhop, A. 2019. Shared measures for evaluating common outcomes of informal STEM education experiences. In A. C. Fu, A. Kannan, & R. J. Shavelson (Eds.), Evaluation in Informal Science, Technology, Engineering, and Mathematics Education. New Directions for Evaluation, 161, 59–86.

Hamlyn, R., Hanson, T., Malam, S., Man, C. Smith, K. and L. Williams. 2020. "Young People's Views on Science Education: Science Education Tracker 2019 Wave 2 Research Report." Available at: https://wellcome.org/sites/default/files/science-education-tracker-2019.pdf [Accessed 10 March 2022].

Hope-Stone Research. 2018. Connect Physics 2017-18 evaluation report. Technical report, SEPnet. Available at: www.sepnet.ac.uk/wp-content/uploads/2018/09/Connect-Physics-Complete-Evaluation-Report-2017-18.pdf [Accessed 23 March 2020].

Husain, F., Wishart, R., Attygalle, K., Averill, P., Ilic, N., and Mayer, M. 2019. CREST Silver Evaluation Report. London: NatCen Social Research and Education Endowment Foundation

Hussar, K., Schwartz, S., Boiselle, E., and Noam, G. 2008. Toward a systematic evidence-base for science in out-of-school time: The role of assessment. Cambridge, MA: Program in Education, Afterschool & Resiliency.

Institute of Education Sciences [IES], and National Science Foundation [NSF]. 2013. Common Guidelines for Education Research and Development. Available at: ies.ed.gov/pdf/CommonGuidelines.pdf [Accessed 23 March 2020].

Laing, K. 2022. 'The contribution of a 'synergic theory of change' approach to democratising evaluation'. Research for All, 6 (1), 8, 1–17.

Kaszynska, P. 2017. The Cultural Value Scoping Project. London: AHRC, PHF and KCL.

Konstantinidi-Sofrona, D. 2021. Evaluation of computing workshops at the science museum. Connected Science Learning 3 (2). Available at: https://www.nsta.org/connected-science-learning/connected-science-learning-march-april-2021/evaluation-computing [Accessed 23 March 2020].

Lloyd, R., Neilson, R., King, S. and Dyball, M. 2012. Review of Informal Science Learning. London: Wellcome Trust.

Mulgan G. 2016. What's wrong with theories of change? Alliance for Useful Evidence Blog, 6th Sept 2016. Available: http://www.nesta.org.uk/blog/whats-wrong-theories-change [Accessed 21 March 2020].

National Research Council. 2009. Learning science in informal environments: People, places, and pursuits. Washington, DC: The National Academies Press. Available at: http://www.nap.edu/catalog.php?record_id=12190. [Accessed 12 March 2020]

NCCPE. 2019. Public Engagement with STEM: Staff and Volunteers Survey. Data summary. NCCPE Available at: https://www.publicengagement.ac.uk/sites/default/files/publication/public_engagement_with_stem_survey_results.pdf [Accessed 23 March 2020].

Noam, G., and Shah, A. M. 2013. Game-changers and the assessment predicament in afterschool science. Cambridge, MA: Program in Education, Afterschool, and Resiliency

Robinson, D. and Salvestrini, V. 2020. The Impact of Interventions for Widening Access to Higher Education: A review of the evidence. TASO report. Available at: https://epi.org.uk/wp-content/uploads/2020/01/Widening_participation-review_EPI-TASO_2020.pdf [Accessed 3 March 2020]

Roy, P., McCrone, T., Rennie, C., Lucas, M., Fletcher, L., Styles, B. and Sims, D. 2021. Generation STEM Work Experience Evaluation Report. London: National Foundation for Education Research (NFER) and Education Endowment Foundation (EEF).

Straw, S., Bamford, S. and Styles, B. 2017. Randomised Controlled Trial and Process Evaluation of Code Clubs. Slough: NFER.

STFC (Science and Technology Facilities Council) 2017. Public Engagement Evaluation Framework. Available at: https://stfc.ukri.org/files/corporate-publications/public-engagement-evaluation-framework/ [Accessed 15 March 2020].

Science Museum Group. 2021. Science capital in practice. Foundations for the future. London: Science Museum Group.

Shimwell, J., DeWitt, J., Davenport, C., Padwick, A., Sanderson, J. and Strachan, R. 2021. Scientist of the week: evaluating effects of a teacher led STEM intervention to reduce stereotypical views of scientists in young children, Research in Science and Technological Education.

Stock Jones, Annable, Billingham and MacDonald. 2016. "Quantifying CREST: what impact does the Silver CREST Award have on science scores and STEM subject selection?" A Pro Bono Economics research report for the British Science Association.

Taplin, D.H., Clark, H., Collins, E. and Colby, D.C. 2013. Theory of change. Technical papers: a series of papers to support development of theories of change based on practice in the field. New York, NY, USA: ActKnowledge.

Terry, M. 2013. Evaluation of the Science Bursaries for Schools and Colleges Programme Final report. Cloud-Champer.

Thatcher, B. and Silversides, K. 2019. Evaluation of the Polar Explorer Programme. York: QaResearch

Thomas, L., Meakin, C., and MacRae, H. 2022. Mission X Evaluation Report. Stirling: Ondata Research.

Valters, C. 2015. Theories of Change. Time for a radical approach to learning in development. London: Overseas Development Institute.

YESTEM Project UK Team. 2020. The Equity Compass: A Tool for supporting socially just practice. Available at: www.yestem.org [Accessed 15 March 2022].

Annex 1: Keywords

Table 2: Keywords used for relevant literature search

Keywords related to evaluation	Participants	Keywords related to engagement	Settings and formats of engagement	STEM discipline
Evaluation Assessment Impact Benefits Challenges Measuring impact Effect Theory of Change	Young people (5–19) Youth Children Adolescents Kids Students Pupils Boys Girls Teens Teenagers Junior Juvenile Minor	Engagement Involvement Participation	Natural History Museums Science festivals Lessons Extracurricular Science clubs Science communication In school Science classes Science centre visit Watching science TV Community coding club Tinkering at home Science museum visit Summer school Outreach Online Informal Science Learning	STEM Science Technology Engineering Mathematics Maths Biology Physics Chemistry Computer science

Annex 2: Interviewees

Role	Organisation
Audience and Insight Manager	Museum
Independent consultant	Public engagement consultancy
Associate Professor in Museum Studies	University
Professor	University
Head of Public engagement	University
Director of Policy	Network organisation
CEO	Network organisation
Co-director	Network organisation
CEO	Independent evaluation organisation
Senior Trials Manager	Independent research Organisation
Director of Research	Funder
Director of Research	What Works Centre
Staff member	US NSF-funded resource centre
Programme Director	US Funder

Annex 3: Interview guides

Interview guide for stakeholders involved in evaluating/researching youth involvement with STEM

UK Research and Innovation (UKRI) is the national funding agency investing in science and research in the UK and brings together the 7 Research Councils, Innovate UK and Research England. UKRI is exploring how funders and policy-makers can drive coordination and improvements in measuring the impact of youth engagement with STEM (ages 5–19). A desk-based research and interviews with experts have been employed to understand the current landscape of evaluating youth engagement with STEM and learning from practices from other sectors.

As an expert in your field, we would like to hear your views on how to best support evaluation practices in youth engagement with STEM.

Your views will be anonymised and included in a report for UKRI.

The current landscape of evaluating youth involvement with STEM

- How would you describe your work in evaluating/researching youth involvement with STEM?
- How would you describe youth engagement with STEM evaluation practice in the UK (in and out of the school)?
- What are the successes and the challenges? How far has the sector progressed in the last 10 years?
- What has worked so far in improving evaluation in youth engagement with STEM in the UK?
- Where are the gaps?
- What are the key frameworks being used in evaluating youth engagement with STEM?
- From your experience, to what extent do STEM organisations have a clear understanding of what they are trying to achieve? E.g. do they have a ToC that they use to evaluate their activities?
- What is the role of RCTs in the evaluation of youth engagement with STEM?
- What is your view on longitudinal studies capturing the impact of youth engagement with STEM? Are there any good examples?
- To what extent and how do you think the evaluation findings of youth engagement with STEM are being used? Do you have any good examples? If they are not used why not?
- What have been the challenges and opportunities that Covid 19 has created for evaluating youth engagement with STEM? What needs to happen next?
- Who is currently conducting evaluations of youth engagement with STEM? Where are the strengths and the gaps in the sector? What skills might be missing?
- Who are some key influencers in evaluating youth engagement with STEM in the UK and internationally?
- What are some key publications, reports, papers on evaluating youth engagement with STEM?

The future of evaluating youth involvement with STEM

- What is needed to develop capacity in evaluating youth engagement with STEM?
- In 10 years, what could youth engagement with STEM evaluation look like?
- What needs to change to achieve this?
- If you were a funder, what would be the key initiatives you would support to strengthen the evaluation of youth engagement with STEM in the UK?
- What is the role of a funder/policy-maker in improving the evaluation of youth engagement with STEM?
- What should a funder/policy-maker prioritise?
- Who should funders/policy-makers be working with to improve evaluation of youth engagement with STEM?
- If you were to conduct a study on improving evaluation in youth engagement with STEM, what would be the key questions to ask?
- What key resources need to be produced to support improving the evaluation of youth involvement with STEM?
- Who needs to be involved in improving evaluation practices of youth engagement with STEM?
- How could they be involved?
- Who else would you recommend consulting as part of scoping how to improve the evaluation of youth engagement with STEM?

Interview guide for professionals involved in evaluating engagement from other fields (e.g. arts, youth sector etc.)

UK Research and Innovation (UKRI) is the national funding agency investing in science and research in the UK and brings together the 7 Research Councils, Innovate UK and Research England. UKRI is exploring how funders and policy-makers can drive coordination and improvements in measuring the impact of youth engagement with STEM (ages 5–19). A desk-based research and interviews with experts have been employed to understand the current landscape of evaluating youth engagement with STEM and learning from practices from other sectors.

As an expert in your field, we would like to hear your views on how to best support evaluation practices in youth engagement with STEM.

Your views will be anonymised and included in a report for UKRI.

The current landscape of impact and evaluation in the youth/arts sector

- How would you describe the evaluation and impact practice in your field?
- What are the successes and the challenges?
- What has worked so far in improving the evaluation in your field?
- Where are the gaps?
- What are the key frameworks being used in evaluating engagement in your field?
- What are your views on creating a common outcomes framework for evaluating engagement in your field? To what extent is this feasible?
- From your experience, to what extent do organisations in your field have a clear understanding of what they are trying to achieve? E.g. do they have a Theory of Change that they use to evaluate their activities?
- What is your view on longitudinal studies capturing the impact of engagement in your field? Are there any good examples?
- What is your view on the usefulness of RCT's capturing engagement impact in your field?
- To what extent and how do you think the evaluation findings are being used in your field? Do you have any good examples? If they are not used, why not?
- What have been the challenges and opportunities that Covid 19 has created for evaluating youth engagement? What needs to happen next?
- Who are the stakeholders/key influencers evaluating the impact of youth engagement?
- What are some key publications, reports, papers on evaluating youth engagement, that have been influential?
- What are some key learnings from improving evaluation in your field that may be useful for the STEM sector?

The future of evaluating youth involvement with STEM

- Based on your experience in your field, if you were a funder, what would be the key initiatives you would support to strengthen the evaluation of youth engagement with STEM in the UK?
- What is the role of a funder/policy-maker in improving the evaluation of youth engagement with STEM?
- What should a funder/policy-maker prioritise?
- Who should funders/policy-makers be working with to overcome the issues in evaluating youth engagement with STEM?
- If you were to conduct a study on improving evaluation in youth engagement with STEM, what would be the key questions to ask?
- Who needs to be involved in improving evaluation practices of youth engagement with STEM?
- How could they be involved?
- Who else from your field would you recommend consulting as part of scoping how to improve the evaluation of youth engagement with STEM?

Annex 4: Table of evaluation reports and papers reviewed

Authors, Year, Type of evaluator	Type of intervention	Type of evaluation	Quasi-experimental, experimental, non experimental	Longitudinal	ToC	Methods used	Limitations	Reporting of things that didn't work	Instruments included in the report	Analysis of student background
Denny, 2021, Consultancy	STEM summer camps	Impact evaluation	Quasi-experimental	No	Yes	Attainment data (mock exams), pre and post surveys, semi structured interviews	Outcome on mental health not measured appropriately.	Yes to some extent	No	No
Thomas, Meaking and MacRae, 2022, Consultancy	STEM enrichment challenge	Process and impact evaluation	Non-experimental	No	No	Pre- post-survey, interviews	Only got feedback from the team leaders not the actual participants	Yes to some extent	Some	No
Archer, M., 2016, in house academic	Pilot of Research in schools	Impact and process evaluation	Quasi-experimental	No	No	Pre- and Post- surveys, feedback forms	Pilot-small sample	Yes	No	% Free School Meals (FSM) at school level
Terry, 2013, Consultancy	Bursaries for placements	Impact and process evaluation	Quasi-experimental	No	No	Pre- and post- surveys, interviews, focus group, observations	The summary report doesn't fully explain the evaluation design. Relying on self-reporting.	Did report that the project had limited impact e.g. in terms of skills.	No	No
CFE Research, Consultancy	Online research project with expert support	Impact and process evaluation	Quasi-experimental	Yes, part of longitudinal work	No	Pre and post surveys, interviews, focus groups	Possible halo effect because of the novelty of the intervention, it was not possible to account for all student and programme characteristics in the comparative analysis.	Yes, great reporting	No. But the regression analysis is included	Ethnicity, gender, families with no prior experience of higher education

Continued on next page

Cilauro and Paull 2019, Consultancy	Summer work placement on research	Process and impact evaluation. Longitudinal, over seven years, qualitative by CFEresearch and quantitative undertaken by Frontier Economics	Quasi-experimental	Longitudinal	No	Pre- post- survey, interviews, focus groups, A level achievement from the NPD data and HE enrolment in the UK from HESA (Higher Education Statistics Agency)	Analysis using NPD data is restricted to pupils in England because of the limited coverage of NPD data. Issues with using the unsuccessful applicants as comparison group: sample size relatively small and successful applicants may differ from unsuccessful due to selection process.	They reported when there was no evidence of impact. More focus on what worked well.	No	Ethnicity, gender, eligible for FSM, income deprivation index
Emembolu et al., 2020, In house Academics	1-hour intervention in schools on STEM careers	Exploratory action research study.	Quasi-experimental	Longitudinal	Yes	Survey, baseline data and follow up data using the 'STEM Career Knowledge and Aspirations Tool' (Padwick, Dele-Ajayi, Davenport, & Strachan, 2016)	The study does not account for other external factors that could influence children's knowledge and preferences. Although interventions were targeted at teachers and families, changes in attitude in these groups were not measured. Cannot draw causal relationship because of time-delay between intervention and children's career choice.	No	Yes	Gender, age, % FSM at school level
Archer M. and DeWitt, 2021, Academics/consultant	School research projects	Impact evaluation	Non-experimental	Longitudinal	No	Post- survey	Difficulty with getting young people complete longitudinal evaluation survey.	Yes	Yes	No
Archer et al., 2021b, Academics	Research projects in schools	Process evaluation	Non-experimental	No	Yes	Post- Survey	Rely on self-reporting. Small survey from a group already bought-in to schools engagement. Results likely less positive from a wider and more representative sample.	Limited	Yes	No

Continued on next page

Shimwell et al., 2021, Academics	Teacher-led STEM intervention, 'Scientist of the Week'.	Impact evaluation	Non-experimental	Longitudinal	No	Open question	Very limited study. Need to explore the efficacy of the method used. Limitation of interpreting a child's meaning from a single word. Need for additional data e.g. from focus group. Lack of comparison group. Need for analysis of effects of repeated interventions over a number of years with a single group.	No	Yes	At school level rather than individuals
Roy et al., 2021, NFER/research organisation	STEM work experience	Impact and implementation and process evaluations.	Experimental	Longitudinal	Yes	Pre- and post- surveys, interviews, observations, Post-16 Learning Aims (PLAMS data held on NPD), Key Stage 2 assessments in 2014, Key Stage 4, 2019 (NPD)	Low implementation fidelity. Lack of impact that was found might be due to issues with programme design or with its implementation. Reported positive outcomes for disadvantaged students but findings based on small sub-group. The 'business as usual' was not established properly.	Yes	Yes	FSM %
Straw, Bamford, and Styles, 2017, NFER	After school Code Club	Impact and process evaluation	Experimental	No	No	Pupil Attitude Survey, Bebras Computational Thinking Assessment, online quiz,. coding quiz, interviews, feedback proforma	Some control group children may have been exposed to resources and approaches used in the intervention.	Yes	Yes	Gender, level of attainment
British Science Association, 2018, Not disclosed	Science Festival	Impact evaluation	Non-experimental	No	Yes	Survey, interviews, Media impact and broadcast coverage	Self reported. Not representative sample.	No	No	Age, gender, ethnicity, postcode
Gammon, 2012, Consultant	Practical biology workshops run by science museums and science centres	Impact and process evaluation	Non-experimental	No	No	Survey, interviews	Mainly self-reported data.	Reported challenges.	Yes	No

Continued on next page

Stock Jones et al., 2016, Consultancy	CREST awards practical science	Impact evaluation	Quasi-experimental	No	No	Propensity Score Matching to create a control group of students, GCSE attainment and AS level study choice from National Pupil Database	Some factors could not be controlled and may have caused upward bias e.g. teacher quality, and enthusiasm for science; parental enthusiasm for science; student enthusiasm for science and whether or not students had also participated in other CREST Awards.	No	n/a	Gender, age, ethnicity, abilities, eligible for FSM
Konstantinidi-Sofrona, 2021, Audience Research Team of the Science Museum Group	Coding workshops for families at science museum	Process evaluation	Non-experimental	No	No	Interviews, observations,	Limited sample of participants. Self-reported data	No	Yes	No
Thatcher and Silversides, 2019, Consultancy	STEM enrichment programme for schools	Process and impact evaluation.	Non-experimental	No	No	Pre- and post- survey, interviews	Not clearly described sources of data. No detailed information about pre- and post- pupil survey that was referenced. Most of the data based on qualitative interviews of teachers reporting on pupil's experiences.	Very limited. Mainly in the recommendations as things to do differently	No	No
Husain et al., 2019, Independent research organisation	CREST awards practical science	Impact and process evaluation	Experimental	No	Yes	Pre- and Post- survey, observations of CREST lessons, interviews, survey, Progress Test in Science (PTS), level 14 (GL Assessment).	Moderate to low security rating. Greater attrition than expected. A self-selection element of recruitment affected the trial's external validity testing. The timing of the follow-up outcome testing was before many students had submitted their projects. Fidelity issues due to the multiple delivery models.	Yes, reported challenges	Yes	Age, gender, eligible for FSM, deprivation index, SEN,