International Journal of STEM Education

RESEARCH

Open Access

Evaluating a complex and sustained STEM engagement programme through the lens of science capital: insights from Northeast England

Annie Padwick[1*], Opeyemi Dele-Ajayi[1], Carol Davenport[1] and Rebecca Strachan[1]

Abstract

Background STEM education providers increasingly use complex intervention models to redress persistent underrepresentation in STEM sectors. These intervention models require robust evaluation to determine their effectiveness. The study examines a complex, sustained intervention intended to build science capital in young people aged 11–15 over 3 years, which drew on science capital theory and related research to inform intervention design and evaluation. When evaluation results differed from those anticipated, process evaluation supported authors to interpret these findings. By outlining challenges faced in the evaluation of a complex, sustained STEM outreach intervention, this paper addresses critique that outreach programmes focus too often on short-term and positive findings.

Results Intervention outcomes were assessed using a quantitative questionnaire adapted from science capital research, issued to pupils at the intervention's baseline (2015), midpoint (2017) and endpoint (2019). Adopting a cohort-based model, the 2015 questionnaire collected a baseline for the Year 7 intervention group (children aged 11–12, $N=464$), and established baseline comparator groups for Year 9 (children aged 13–14, $N=556$) and Year 11 (children aged 15–16, $N=342$). The Year 7 intervention group was re-evaluated again in 2017 when in Year 9 ($N=556$), and in 2019 when in Year 11 ($N=349$). Analysis explored differences in science capital between the intervention and comparator groups and identified lower composite science capital scores and greater proportions of low- and medium-science capital in the intervention group when compared with the two comparator groups. A rationale for this emerged from the subsequent process evaluation.

Conclusions This study's main contribution is the provision of nuanced insight into the evaluation of STEM interventions for use by others evaluating in similar circumstances, particularly those adopting sustained or complex delivery models. This paper concludes that assessing the effectiveness of complex interventions cannot rely on quantitative evaluation of outcomes alone. Process evaluation can complement quantitative instruments and aid interventions to better understand variability and interpret results. While this study highlights the value of science capital when designing intervention models, it also illustrates the inherent challenges of using an outcome measure of 'building science capital', and quantifying levels over an intervention's course.

Keywords STEM education, Science capital, Complex intervention, Evaluation research, Implementation theory, Process evaluation

*Correspondence:
Annie Padwick
annie.padwick@northumbria.ac.uk
Full list of author information is available at the end of the article

Introduction

Many STEM (science, technology, engineering, and maths) outreach programmes in universities share a common purpose of increasing the overall participation of students in STEM (Sadler et al., 2018). This shared goal often feeds into broader, more long-term commitments to address the enduring challenges of workforce shortages (Dromey, 2021; Neave et al., 2018) and under-representation of certain groups in the STEM sectors (APPG, 2020). With their long history of social and civic engagement, universities tend to rationalise their STEM engagement on several grounds (Sadler et al., 2018). Economic arguments are focused on the need to meet the growing demand for greater numbers of workers for the STEM industries (Neave et al., 2018). Rights-based approaches argue that inequities in STEM also give rise to discrepancies in opportunity and outcome between privileged and disadvantaged groups, while a justice-orientated framing finds the lack of an extension of rights to legitimate participation in STEM disciplines a fundamental injustice (Calabrese-Barton & Tan, 2020).

STEM outreach teams are aided by STEM education researchers who are investigating, understanding, and defining the nature of under-representation (Archer et al., 2015; Calabrese-Barton & Tan, 2010; Carlone & Johnson, 2007). Science capital research has offered insights to STEM outreach groups by providing theoretical underpinnings for patterns seen in young people's participation in science. Developed by the ASPIRES team, this conceptualisation illuminates the different types of economic, social, and cultural capital that relate specifically to science, particularly those with potential to leverage support and enhance a person's attainment, engagement, and participation (Archer et al., 2015). Science capital researchers are also exploring how science capital can be used to inform the design and development of STEM education interventions and support the practical evaluation of their outcomes (DeWitt et al., 2016). This endeavour has been continued by others (Jones et al., 2022).

There are many STEM education providers ready to apply new theories, put recommendations into practice and test them out in their own contexts (Powell et al., 2018), however, applying research theory to intervention theory requires STEM practitioners to first work through a number of design considerations (Kezar, 2011; Reinholz et al., 2021). Among others, the choice of which theory to apply to shape their practice (Powell et al., 2018; Reinholz et al., 2021), how to translate and shape the theory into a practical intervention model, and then last which evaluation designs, methods, and instruments to use to determine if their endeavours are successful (Boaz et al., 2021; Crawford et al., 2017; Sarmiento-Marquez et al., 2023).

Solutions to the problem of under-representation require greater understanding, both about what does and does not 'work' in STEM education and outreach, and for whom this is the case (McKinnon, 2022). The evaluation of STEM engagements, however, is challenged by several factors including frequent under-funding (Wilkerson & Haden, 2014), lack of strategic planning, targeting and poorly designed outcome measures (Sadler et al., 2018), which individually or collectively cause variability in the intensity and quality of evaluations (McKinnon, 2022; Ziegler et al., 2021). Furthermore, longer term evaluation models are required now that STEM engagement providers are moving to develop deeper, more impactful engagements, by designing more complex interventions that take place over longer time periods (Archer et al., 2021; Sadler et al., 2018). Testing the effectiveness of such interventions requires a greater understanding of appropriate evaluation approaches and methods (Archer et al., 2021; Boaz et al., 2021; Outhwaite et al., 2020). However, the evaluation of outreach frequently falls to project delivery staff, with institutions often providing little in the way of tangible support (Crawford et al., 2017). Meanwhile, many STEM outreach and education providers fail to report evaluations publicly (Banerjee, 2016; Rosicka, 2016), and those that do habitually frame evaluations within success story narratives (Ziegler et al., 2014). Yet despite many interventions claiming successes in their outcomes to increase participation in STEM (Biesta, 2010), numbers of students participating in STEM remain stubbornly low (APPG, 2020; Reed-Rhoads, 2011).

Although outcome evaluation provides understanding about the progress made toward achieving target outcomes and objectives, it is only the analysis of an intervention's implementation that enables better understanding of how and why it has worked, or has not worked as intended (Humphrey et al., 2016; Outhwaite et al., 2020). This is particularly the case with complex interventions taking place in the real-world (Reynolds et al., 2014). Process evaluation, a method stemming from implementation theory, uses a systematic approach to documenting and accounting for deviations from the intervention as intended, by reporting the actual implementation, take up and context for an intervention (Outhwaite et al., 2020; Reynolds et al., 2014).

This case study paper presents the results from the evaluation of a complex STEM outreach intervention involving 15 secondary schools[1] in Northeast England and multiple stakeholders (children and young people,

[1] In England, primary schools cater for pupils between the ages of 5 and 11 and secondary schools for pupils between the ages of 11 and either 16 or 18.

teachers, and parents) engaged over a 3-year sustained period. The host university received funding to establish an outreach group and work with a partnership of schools to explore provision of sustained STEM support within existing school cultures and structures. The intervention's aims were to build science capital in children and young people in participating schools and develop a blueprint for improving the uptake of STEM subjects that could be used by universities in other regions. The evaluation context was one which examined the effectiveness of the outreach intervention within its complex, real-world and dynamic delivery environment, rather than in a controlled efficacy trial.

Established in 2014, the intervention drew on the, then nascent, conceptualisation of science capital to develop its Theory of Change (ToC) (Davenport et al., 2020). Science capital enabled the authors to identify the elements most important to young people's participation in science, and those that might potentially have the most lasting impact. The group developed a flexible offer to partner schools, with activities to be selected by schools and co-delivered with the project team. The authors were also keen to explore how elements of young people's science capital might change because of the sustained STEM education programme. Encouraged by ASPIRES research (De Witt et al., 2016), the expectation was that the intervention would increase levels of science capital. To aid in this, some of the quantitative instruments used within science capital research were adapted to be used as part of the outcome evaluation of the intervention in three partner schools.

This study first examines data produced by the quantitative science capital inspired instruments in the outcome evaluation, before going on to use a process evaluation framework to consider the complex sustained intervention through seven factors affecting implementation (Humphrey et al., 2016). Data from these evaluations are then drawn together to explore the following research questions:

> *What are the affordances of science capital in the development of outcome measures for STEM interventions?; To what extent can a quantitative index provide sufficient information to evaluate a complex STEM intervention?; and How appropriate was the chosen evaluation model for the evaluation of a complex sustained intervention with a flexible delivery model?.*

After nearly 10 years of development, science capital has become well-embedded in the STEM education landscape, being used in intervention and outcome design of several STEM interventions (Bryan et al., 2022; Harris et al., 2018; McCracken, 2019), including the one presented in this paper. In response to the critique that outreach programmes report only short-term and positive findings (Sadler et al., 2018; Ziegler et al, 2021), this study makes explicit the considerations involved in design of the science capital inspired intervention and evaluation and provides reflection on whether the methodological choices were subsequently found to be appropriate and sufficient. The overall intention is to set out the lessons learned for the benefit of other STEM educators and practitioners working with a similar purpose. By articulating both the challenges faced and failed attempts (Ziegler et al., 2021), this paper hopes to evolve the wider evaluation narrative of STEM outreach providers and support the development of a constructive environment, where educators can learn from one another and develop skills and knowledge about what works, based on robust evidence.

Background

Science capital

Understanding the various factors that influence children's participation in science and science careers has engaged researchers for decades (Christidou, 2011; Gardener, 1975; Osbourne et al., 2003). Time and time again, research has found that children and young people enjoy studying science in school, but this interest rarely develops into science-related aspirations (Archer et al., 2012; Osborne et al., 2003). In recent years, researchers have considered how influences outside the school environment, including family background, daily life and circumstances, affect children and young people's participation in science and STEM subjects (Archer et al., 2012; Gokpinar & Reiss, 2016; Moote et al., 2020). Bourdieu's theories of capital, habitus and field have been particularly useful to researchers investigating the persistent inequalities in STEM (Archer et al., 2012; Black & Hernandez-Martinez, 2016; Claussen & Osbourne, 2013; Gokpinar & Reiss, 2016; Nicolaisen et al., 2023).

The ASPIRES study tracked the development of young people's science and career aspirations from age 10 to 14 (from 2009 to 2013), concluding that although students say they learn interesting things in science and think that scientists do valuable work, they do not aspire to science careers (Archer et al, 2012). The study's findings were considered in the context of Bourdieu's theory of social reproduction, particularly the concept of 'capital', which generated the concept of 'science capital'. Rather than describing a new 'type' of capital, Archer et al. asserted that science capital should be seen as:

> *"A conceptual device for collating various types of economic, social and cultural capital that specifi-*

cally relate to science." (Archer et al., 2014 in Archer et al., 2015, pp. 5)".

Science capital provides a theory-based explanation for patterns in young people's participation in science, encompassing a number of different dimensions. These are scientific forms of cultural capital (scientific literacy, scientific-related dispositions/preferences, knowledge of transferability of science skills and qualifications, symbolic knowledge about the transferability of science); science-related behaviours and practices (science media consumption, participation in out-of-school science learning contexts); and science-related forms of social-capital (parental scientific knowledge/qualifications, knowing someone who works in a science job, talking to others about science) (Archer et al., 2015). ASPIRES found that certain components of science capital (science literacy, perceived transferability and utility of science, family influences) seemed more closely related to anticipated future participation and identity in science than others and highlighted the important role families play in shaping students' aspirations and participation in science. The research concluded that where families possess high levels of science capital, young people are more likely to aspire toward a career in science by the age of fourteen than those in families with lower levels of science capital (Archer et al., 2012).

Though originally formed as a theoretical model for interpreting patterns in children and young people's aspirations for STEM, Archer and colleagues also considered how science capital might be used as a guide to measure and evaluate outcomes and impacts from STEM education interventions. Archer et al. state:

"…it may be interesting and useful … for the science education community to be able to 'measure' and determine levels of science capital at scale… to be able to delineate what they are seeking to change through their practice and why and to assess to what extent they have been successful, or not, in these efforts" (Archer et al., 2015 pp. 928).

The ASPIRES study combined quantitative online surveys of a student cohort with longitudinal interviews with a selected sub-sample of students and their parents. A subsequent iteration of the survey, as part of a related project, was later used to refine a measure of science capital and create a shorter, more usable 'science capital index', consisting of 14 items, which was used to generate an overall science capital score for individuals (Archer et al., 2015). This index was used to generate numerical scores that were used to divide young people into three groups—those possessing low, medium, and high levels of science capital. In a national survey, 5% of the study sample were classified within high, 68% in medium and 27% in low science capital groupings (De Witt et al., 2016).

Implementation theory

Implementation theory draws on the disciplines of psychology, sociology and organisational theory, as well as models and frameworks that have emerged from implementation science (Nilsen, 2020). Application of this theory has its foundation in health-care research, though usage in educational intervention contexts is gaining ground (Outhwaite et al., 2020). Implementation theory offers theories, models and frameworks that can be used to gain insight into how an intervention works and under what circumstances, and the mechanisms by which the implementation is more likely to succeed (Nilsen, 2020). A key aspect involves assessing the variability in intervention implementation across different contexts (Peterson, 2016). Implementation evaluation proponents argue that measuring the effectiveness of an intervention by outcomes alone could lead to misleading results, and that outcomes should always be considered alongside an examination of the implementation of the programme (Humphrey et al., 2016; Nilsen, 2020; Outhwaite et al., 2020).

As a method for assessing an intervention's implementation, process evaluations provide a systematic approach to documenting the intervention as delivered, compared to what had been intended. Implementation can be examined in terms of fidelity, reach and dose delivered, as well as any unanticipated additional activities and adaptations made. Process evaluation in education has predominately focused on fidelity, the extent to which the intervention is delivered as intended, though implementation can also be examined in other areas, such as adaptations and quality (Hoffman et al., 2014; Outhwaite et al., 2020).

Humphrey et al. (2016) has established seven important factors that affect implementation in educational settings (see Table 1). Adherence and quality consider how well the programme was delivered, while dosage, reach and responsiveness are concerned with assessing programme take-up by target audiences, and programme differentiation and monitoring of control/comparison groups consider how well the intervention's impacts can be distinguished from other influences.

By bridging the gap between the intervention as intended and effective practice (Fixsen et al., 2009), process evaluation complements quantitative approaches, as well as holding value for theory-based evaluation approaches. Furthermore, in the case of interventions

Table 1 Dimensions of and factors affecting implementation of an education programme

Factor	Description
Adherence	The extent to which implementers adhere to the intended treatment model
Quality	How well the different components of the intervention are delivered
Responsiveness	The degree to which participants engaged with the interventions
Dosage	How much of the intended intervention has been delivered and/or received
Reach	The reach and scope of participation
Programme differentiation	The extent to which intervention activities can be distinguished from other existing practice
Monitoring of Control/ Comparison groups	Determination of the 'counter-factual', i.e., that which is taking place in the absence of the intervention

Adapted from Humphrey et al., (2016), pg. 6

which appear to fail in realising their outcomes, process evaluation can provide an analysis method for determining whether this was due to poor programme design or poor implementation (Askell-Williams et al., 2013).

Intervention and evaluation design

This section presents an outline of the intervention design and the evaluation design used in this study, and the rationale for these choices.

STEM outreach intervention intended model and theory of change

The authors formulated an initial intervention model and later a Theory of Change (ToC) (Davenport et al., 2020) by drawing on the initial conceptualisation of science capital (Archer et al., 2012, 2013) and other influential reports in the literature prior to 2015 (HEFCE, 2014; Hughes et al., 2013; Institute of Physics, 2013; Murphy & Whitelegg, 2006). The core components of the intended model are described below, with an explanation of the intervention implementation presented in the results section following.

From the literature, the ToC identified the intervention's key components as early engagement with children, followed by sustained intervention over the course of their education journey (pre-school to post-16), combined with a shared vision and collaboration across the study period (Fig. 1). Furthermore, joined up working with the children's key influencers—i.e., their parents, families and communities and teachers and school communities—was expected to be important to success, as was raising awareness about the transferability and utility of science through the focus on increasing young people's knowledge of possible STEM careers (Davenport et al., 2020).

A key driver in development of the intervention model was the understanding that one-off or short-term interventions are unlikely to have lasting impacts on aspirations or future participation in STEM (Archer et al., 2021). The model was thus one of a sustained partnership with a network of schools. This research paper focuses on the design and evaluation of the secondary school aspects of the programme, examining evaluation data from three partner schools. The design and evaluation of the engagement programme with primary schools has been reported elsewhere (Emembolu et al., 2020; Padwick et al., 2016).

In this paper, the authors use 'sustained' to mean long-term engagement over the course of a year or more involving frequent or repeated activities (Archer et al., 2021). It was intended that young people would experience several mutually reinforcing activities, ideally a few times a year, over the course of the intervention. Participating schools could select from available activities including school assemblies, in-class workshops, after-school workshops, STEM clubs, summer schools and careers events. By creating new and broader opportunities to engage with science, the intervention's activity was intended to enrich rather than replace the taught curriculum. This delivery design may be described as a 'loose enabling framework', in that intervention within each school was based on a common approach at a broad level, while the specific components for action in each school were not specified (Humphrey et al., 2010). Teachers were also encouraged to identify gaps in their knowledge and provision and suggest how the intervention could complement existing school activities. The ability for schools to make local adaptations and request additional elements for inclusion in the delivery of the intervention supported their expectations for professional autonomy and flexibility, while encouraging buy-in and enhancing the 'fit' between intervention and school (Education Endowment Foundation, 2019).

The intervention model incorporates working with key influencers of young people (their teachers and their families) to model good practice and implement the ToC. The identified mechanism for change was that influencing the key influencers of young people had the potential to create the greatest impact on young people and a long-term legacy (Davenport et al., 2020). It was thus intended that teachers receive continuing professional development in the core principles, such as integrating STEM careers and mitigating unconscious bias while also being exposed to the principles of engagement during the delivery with young people. Since the intervention model recognises parents and families as key influencers of young people's aspirations for, and participation in science, the intervention also looked at how best to involve parents and carers. Since the opportunity for parent–school relations wanes as young people get older (Deslandes & Cloutier,

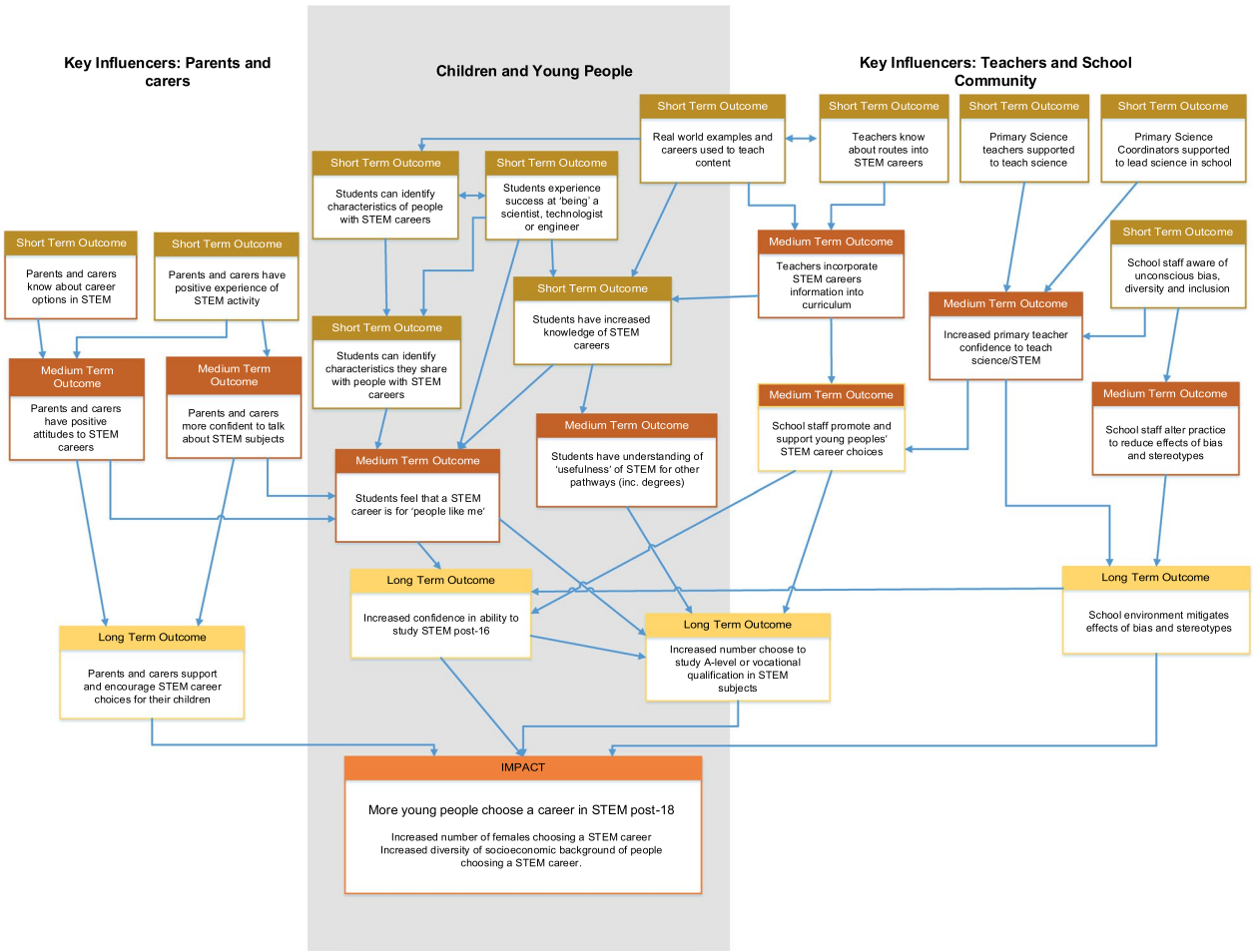

Fig. 1 ToC model (Davenport et al., 2020)

2002), the intervention at secondary level focused on meeting parents at career events and parents' evenings.

Evaluation design considerations

The choice of which evaluation design to choose for a STEM intervention is not always straightforward. Key considerations include the purpose of the evaluation, programme structure and circumstances, delivery model, resource available for evaluation (Rossi et al., 1999), and the expectations of stakeholders, commissioners, and funders.

Initial focus of the intervention's evaluation design was to assess realisation of project outcomes and demonstrate intervention impacts. Aligned with the long-term outcome for '*increased number choose to study A-Level or vocational qualification in STEM subjects*', the initial design sought to track the cohort's post-16 subject choices[2] along with Higher Education destinations via post-intervention analysis using the National Pupil Database and Higher Education Statistics Agency data sets, and to compare this against national and regional averages. However, after undertaking this analysis, the approach was found not to be suitable for small-scale intervention monitoring, due to reporting thresholds, variation, and availability of data (Padwick & Davenport, 2022). This evaluation focuses on addressing the aim to '*build science capital in young people*' by measuring science capital over the course of the intervention. When the outcome evaluation identified no evidence of impact, the authors turned to process evaluation to aid their understanding as to why the intervention did not work as intended.

Choice of evaluation methods and instruments

As the evaluation design sought to measure science capital, the authors chose to utilise methods and instruments

[2] In England, science is compulsory for pupils up to the age of 16. Beyond 16 pupils may choose to continue to study science subjects including physics.

designed for that specific purpose. From the choice of quantitative questionnaire and qualitative interviewing utilised in the early ASPIRES study, the authors decided to adapt the quantitative instrument which was obtained directly from the researchers (Archer et al., 2012, 2015). As this evaluation design work predated the publication of the science capital index (Archer et al., 2015) that instrument itself was not used.

The rationale for using quantitative methods took into account the intervention funder's request to report evaluation findings numerically. In addition, as the potential intervention scale was large, quantitative methods allowed the team to capitalize on these large numbers to assess cohort level changes at scale. The intention was that the instrument would enable changes resulting from the sustained STEM education programme to be quantified. It was anticipated that at each datapoint data would be used to generate both an individual's numerical 'science capital' score and to classify science capital levels into high, medium, and low groupings (Archer et al., 2015), enabling comparisons across the time series. Moreover, a quantitative research instrument provides simplicity and speed of completion for those within the classroom, with relative ease of data processing for the STEM intervention and researchers (Craig, 2014). While qualitative approaches might have complemented these quantitative approaches by providing rich data and greater insights of intervention on individuals, the collection and analysis of such data can be resource intensive (Margoluis et al., 2009). Due to allocation of resources available for the intervention's evaluation, the authors made the decision to exclude the qualitative methods used in ASPIRES (Archer et al., 2012, 2015) from the evaluation design. The inclusion of a process evaluation was not included in the initial evaluation design.

In developing the quantitative instrument, the programme team reviewed the factors contributing to science capital before identifying those most closely aligned to the ToC and intervention outcomes. The four dimensions identified were: science conversations, science social contacts, science self-concept, and utility of science. This led to development of a short questionnaire with seventeen items based around these central themes alongside six demographic items (see Additional file 1). The original data analysis plan included Principal Component Analysis (PCA) of questionnaire data as means of validating these components and reducing dimensionality of the data set from a large number of variables into fewer for analysis. This was undertaken and is reported in Additional file 2.

Since the inclusion of the process evaluation came as a response to the results of the outcome evaluation, it required the use of existing data. Data used were routinely collected monitoring data on activities, audience, reach, and delivery staff involved. The evaluation of the implementation of the intervention focused on the seven factors as outlined by Humphrey (2016).

Evaluation models

An intervention can be described as a 'complex intervention' (Craig, 2008) if it has multiple interacting components acting independently or interdependently (Reynolds et al., 2014). The STEM intervention addressed by this study aligns to this definition by containing several interacting components, targeting several groups and organisations at different levels and using a less prescriptive delivery model or 'loose enabling framework' which in turn leads to a number and variability of outcomes (Craig, 2008). It cannot be taken for granted that complex interventions delivered in 'real-life' dynamic contexts will be implemented exactly as it was envisaged at the design stage (Margoluis et al., 2009; Reynolds et al., 2014).

Experimental randomised control trials (RCT) are not always the most appropriate or ethical fit for the evaluation of real-life education interventions, particularly those which are complex, sustained or dynamic (Connolly et al., 2018; Sullivan, 2011; Wells et al., 2012). This is due to challenges in 'blinding' learners from knowing they are part of an intervention, the contamination effects of students from one group mixing with another within a school setting, and because on ethical grounds, interventions aimed at developing learning should not be withheld from one group over another (Sullivan, 2011). In addition, while some programmes were successful in RCT trials, they have been found to be less successful, or work differently, once applied to real school settings (Askell-Williams et al., 2013).

Interrupted time series is a non-experimental method involving tracking over a long-term period before and after a point of intervention to assess its effects. It can be used to estimate the impacts of interventions when randomised controlled trials and quasi-experimental designs are not feasible (Hudson et al., 2019; St. Clair et al., 2014). For this study, a time series cohort design was selected to consider the cumulative impact of a range of different activities over the intervention's duration (Margoluis et al., 2009). Data were collected from children in alternative years of their secondary schooling: year 7 (age 11–12), year 9 (age 13–14) and year 11 (age 15–16).[3] The choice to sample from alternating school years and draw from a smaller number of evaluation

[3] In the English school system, year groups are numbered from the point at which pupils start compulsory schooling at age 5. Thus, pupils at the start of secondary school (age 11–12) are said to be in Year 7, pupils in Year 9 are aged 13–14, and pupils in Year 11 are aged 15–16.

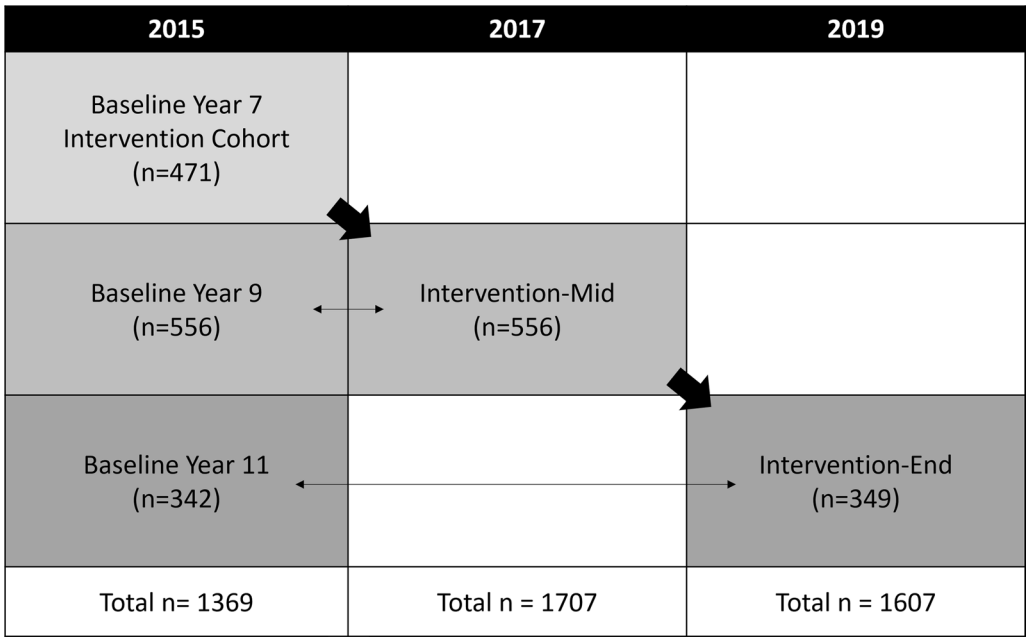

Fig. 2 Diagrammatic representation of intervention and age-aligned comparator cohorts drawn from baseline

schools reflected the time and resources available for evaluation (Fig. 2).

As the strength of time series cohort design can be improved by inclusion of a comparison group (Sullivan, 2011), the study follows an *age-period* cohort design (Yang & Land, 2013) by including comparison cohort groups to mitigate the effects of the development within young people (*age*) concerned over the intervention's 3 years (*period*). As well as providing the baseline for the intervention group (Baseline Year 7), the baseline data gathered in 2015 provides two comparator groups to assess cohort changes over time. The 'Baseline Year 9' are those pupils in Year 9 in 2015 not exposed to the intervention but in an equivalent year group in terms of age range. These act as comparator group to the intervention cohort at the midpoint 'Intervention-mid', that is to say those entering Year 9 in 2017 who have now had 2 years of intervention. The 'Baseline Year 11' are those pupils in Year 11 in 2015 not exposed to the intervention, who are in an equivalent year group in terms of age range and act as a comparator group to the intervention cohort at the endpoint 'Intervention-end', thus those entering Year 11 in 2019 who have now had three (academic) years of intervention. The intervention and comparator cohorts are shown in Fig. 2. The intervention cohort over time is indicated by the bold arrows—pupils in Year 7 at the start of the programme (2015) and in Year 11 at the end of the programme (2019) with their aligned comparator groups for each age period indicated by the thin arrows. As the comparator groups are drawn from the baseline year of evaluation, this does not constitute a matched sample. However, pupils in each school in Year 7 in 2017 and 2019 were also considered to ensure the incoming student cohort had similar characteristics to the intervention cohort. The baseline comparator cohorts were considered adequate as they had been drawn from the same schools and had similar characteristic sets, including in gender split and educational experiences.

Methods

15 secondary schools were recruited to the project in 2015, drawn from a potential pool of 77 secondary schools in 5 participating Local Authorities. School recruitment choices were made on recommendation from local authority education officers to target schools with a high percentage of children on free school meals. From these 15, four secondary schools were selected to act as evaluation schools from which data would be collected from young people. Selection ensured geographical representation from each of the participating local authorities. At the end of the project, only three evaluation schools returned data across all three collection points, and thus it is these schools that form the sample for this study.

The research sample was drawn from young people in Years 7, 9 and 11 in these three schools. The questionnaire was issued by teachers either during registration or science lessons. Not all pupils completed the questionnaire, and schools were not asked to indicate what proportion of their enrolled pupils took part in the data

collection. A baseline was established during September–December 2015, with further data collections in September–December 2017 and 2019. This resulted in 1369 responses to the baseline questionnaire, 1707 to the midpoint and 1607 to the endpoint questionnaires, representing between 55% and 70% of possible responses per school. Figure 2 breaks down returned responses for the intervention group in 2015, 2017 and 2019 and the two age-aligned comparator cohorts (Baseline Year 9 and Baseline Year 11).

Data analysis

Likert scale items were coded as follows: strongly disagree = 1, disagree = 2, neither = 3, agree = 4, and strongly agree = 5, with negatively framed questions reverse coded (see Additional file 1 for details). 'When not in school, how often do you talk about science with other people?' was coded similarly, e.g., 'how often' as 1 = never and 5 = always. 'Who do you talk to most about science?' was rescaled according to the value within science capital conceptualisation, therefore, family = 5, friends = 4, classmates = 3, teachers = 2, other = 1, no one = 0. Data analysis was conducted in IBM SPSS Statistics Data Editor (26).

A composite 'science capital score' was created which summed all seventeen items and generated a possible score range of 17–85. Cronbach's Alpha was used to test the internal consistency of the science capital composite and assess whether the Likert scale items measured were the same general construct. Descriptive statistics were obtained for a science capital score of both intervention and aligned comparator cohorts.

Following the methodology described by Archer et al. (2015), the science capital score was used to assign participants into three groups, with the score subsequently divided into thirds, with low science capital assigned to the first third (17–39), medium science capital the next (scores of 40–62) and high science capital the remaining third (scores of 63–85). The descriptive statistics obtained for high, medium, and low science capital groupings for intervention and comparator cohorts determined the differences between groups. Statistically significant differences between the intervention group and baseline comparator groups were examined using Welch's T test. p values of < 0.05 were considered statistically significant.

For the process evaluation, the engagement data spreadsheet for the whole intervention was examined before extracting data for the three evaluation schools. Activities were categorised into assemblies; careers events; class workshops; clubs and informal activities; participation in events offered by partner organisations; and teacher CPD. For each delivery year frequency counts were obtained for activity type, duration of activities, and the numbers of children and teachers engaged. The analysis also examined activity take up by year group. Given that intervention activities were offered to year groups across the whole school, the analysis next considered specifically which of the activities the intervention cohort received. Process evaluation findings were discussed with members of the outreach delivery team and reviewed in line with seven implementation factors (Humphrey et al., 2016).

Results

Outcome evaluation results

The Cronbach's Alpha test for the composite science capital score produced values of 0.8, which indicated very good internal consistency levels for the items.

The composite science capital score for the intervention cohort at baseline (Year 7) was $M = 58.1$ and SD = 8.7. The composite science capital score for the intervention at midpoint ($M = 56.8$, SD = 8.9) and intervention at endpoint ($M = 55.9$, SD = 8.4) were both lower than the baseline score.

These scores at intervention-mid and intervention-end were also examined alongside their age-aligned comparator cohorts. Welch's T test results for the intervention-mid indicate that the composite capital score was significantly higher in the Baseline Year 9 comparator ($M = 57.9$, SD = 7.1) than in the intervention cohort at midpoint ($M = 56.8$, SD = 8.9), $t(1051) = 2.4$, $p = 0.017$. Results for the intervention-end indicate no significant difference between group means in the Baseline Year 11 comparator ($M = 55.6$, SD = 8.0) and the intervention cohort at endpoint ($M = 55.9$, SD = 8.4), $t(676) = -0.478$, $p = 0.316$) (Fig. 3).

The proportion of respondents with high, medium, and low science capital were determined at each timepoint for both intervention group and their age-aligned baseline comparators (Table 2). The only significant difference was found between the intervention mid and Baseline Year 9 comparator group, with the baseline comparator significantly higher than intervention mid [Chi Square χ^2 $(2, N = 1047) = 21.047$, $p \leq 0.001$].

Process evaluation results

Analysis of engagement data found that within the study period 2015–2019, the evaluation schools engaged with an average of 19 activities. There were 25 engagements during 2015–2016, reducing to 6 during 2016–2017, 15 engagements in 2017–2018 and 12 in 2018–2019. Table 3 shows the variability of numbers of activities per school over time, including the particularly low take up of teacher CPD.

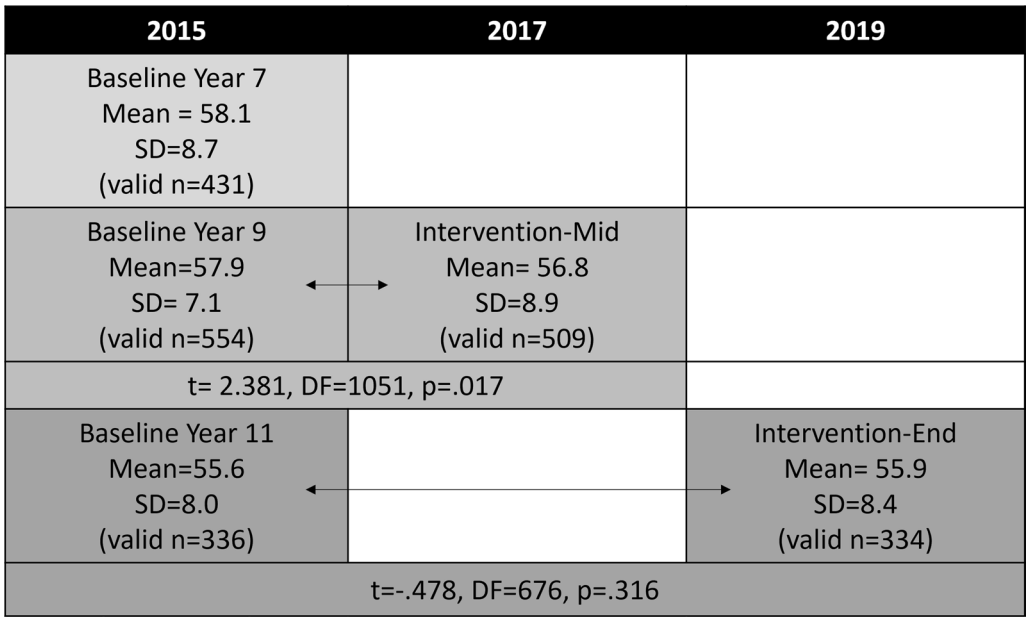

Fig. 3 Composite science capital scores for intervention group and age-aligned comparator groups: descriptive statistics and Welch's *T* test results

Table 2 Low, medium and high science capital groupings at intervention stages with group differences to age-aligned comparator groups by Chi Square Test

	% Low science capital	% Medium science capital	% High science capital	X^2	DF	*p*
Baseline Year 7	1.2	58.1	40.7			
Intervention Cohort at Midpoint (Year 9)	3.6	69.4	27.0	21.047	2	0.001
Baseline Year 9 Comparator	0.6	62.8	36.7			
Intervention Cohort at Endpoint (Year 11)	2.3	79.3	18.4	3.255	2	0.196
Baseline Year 11 Comparator	1.2	75.8	23.0			

A review of the process evaluation data was undertaken by the outreach team against the seven implementation factors (Table 4). The intervention adherence was challenging to measure due to the 'loose enabling framework', meaning there was no intended model to adhere to. While the full programme offer encompassed all the important features of the intervention, not all programme elements were taken up by participating schools. Table 3 shows many gaps in provision as regards certain of the offer's activities. Assemblies and classroom workshops were popular and were able to engage large numbers of pupils; however, whole year or whole class engagements can also be considered some of the lower impact activities offered within the overall programme. The potentially higher impact activities such as continuing professional development for teachers and engagements with parents were undertaken less frequently. Only one school took up the offer of continuing professional development for staff during the intervention period, while engagement with parents at all schools was minimal and sporadic, largely effected through school open evenings and careers events. Thus, although the activities delivered by the outreach team adhered to the intended treatment model, the 'whole programme offer' was not fully adhered to in any of the evaluation schools. Both delivery team assessment and feedback from partner schools judged the quality of the intervention as good, although given available data, it was not possible to explore this conclusion in any more depth.

While some evaluation schools were more responsive to the STEM engagement programme than others, the overall responsiveness was lower than predicted. However, it is worth noting that other participating schools had a higher take up of activities, meaning the three evaluation schools were not representative of many schools within the overall programme. Frequency counts of the activity types shows that overall activity responsiveness and take up of different activities waxed and waned at

Table 3 Uptake of different intervention activity types per year

		2015–2016	2016–2017	2017–2018	2018–2019	Activity total	School Total
School 1	Assemblies	–	–	1	–	1	24
	Careers events	1	–	2	1	4	
	Class workshops	4	–	3	7	14	
	Clubs and Informal	–	1	–	2	3	
	Partner org events	1	1	–	–	2	
	Teacher CPD	–	–	–	–	0	
School 2	Assemblies	6	–	–	–	6	19
	Careers events	1	2	1	–	4	
	Class workshops	3	–	1	–	4	
	Clubs and Informal	–	–	3	–	3	
	Partner org events	1	–	1	–	2	
	Teacher CPD	–	–	-	–	0	
School 3	Assemblies	5	–	-	–	5	15
	Careers events	–	1	-	–	1	
	Class workshops	1	1	1	2	5	
	Clubs and Informal	–	–	–	–	0	
	Partner org events	2	–	1	–	3	
	Teacher CPD	–	–	1	–	1	

Table 4 Examination of the 7 implementation factors of the sustained STEM intervention

Factor	Consideration of factor in this study
Adherence	a) The intervention was delivered to a loose enabling framework, and therefore, measuring adherence is challenging b) The full programme offer was not delivered in any of the evaluation schools or participating schools c) Activities taken up were delivered according to the intended treatment model
Quality	a) Feedback from schools and reflection on quality of intervention were good b) Data used within the process evaluation did not allow for this factor to be examined in much depth
Responsiveness	a) There was lower than expected take up of number of programme activities in all schools b) Levels of engagement waxed and waned over the course of the intervention according to changes in school circumstances, particularly related to the school accountability regime
Dosage	a) Schools generally took up low-impact activities, such as whole-school assemblies and class workshops b) Some of the higher impact elements of the programme offer were not taken up readily (CPD and parental engagement) c) Young people within the measured cohort study group received low levels of intervention at an individual level
Reach	a) Lower than expected reach among some year groups of the schools b) There was greatest reach among low-impact activities, such as whole-year group assemblies c) The take-up of intervention activity among the measured intervention cohort was low
Programme differentiation	a) The intervention was only one of many things happening in young peoples' lives during the intervention period b) New curriculum and qualifications were introduced during this time (2015)
Monitoring of Control/ Comparison groups	a) Comparator groups have been employed, accounting for natural changes as young people age b) Design of comparator group cannot take into account the counter-factual of education policy and other changes occurring during the programme

various point during the intervention at an individual school level. When reflecting on conversations with school leads the delivery team concluded that responsiveness and engagement levels particularly related to school accountability regimes.

Reach of the STEM engagement programme at school level met expectations, in that it engaged with 15 secondary schools. Although STEM interventions were offered universally to the different year groups across the whole school, individualised school selection resulted in lower reach among younger year groups (years 7, 8, and 9). Examination of data from the intervention cohort, i.e., the activities delivered to year 7's in 2015, year 9's in 2017, and year 11's in 2019 leads to the conclusion that the intervention dose in the evaluation schools was very low at an individual level. Of the intervention group in the

most engaged evaluation school, young people engaged in one classroom workshop in 2015–2016, another workshop and a careers fair in 2017–2018, and subsequently a science talk in 2018–2019. In another evaluation school, young people in the study cohort only received activities during the first 2 years and none in the final year. Thus, both at a cohort and individual level, the intervention dose and reach proved to be insufficient to generate the intended outcomes in these schools.

Due to the study's longitudinal nature, comparator cohorts were employed to mitigate natural development and change in young people over time as they age. However, a model that uses comparator cohorts drawn from the baseline disallows the determination of the temporal counter-factual, i.e., that which is taking place during the study period but in the absence of an intervention. Possible influential and often inter-related factors (Blickenstaff, 2005) can be found in social context, school context, social environment and student characteristics. An example of this is the introduction of a new curriculum and changes to qualifications that occurred within the period of study (Long, 2017), which both negatively affected teacher workload and young people's attitudes to and engagement with subjects under study (Neumann et al., 2016). However, the data available from the outcome evaluation or process evaluation disallowed any assessment of other influences on the intervention group during this period.

Discussion

This paper presents results from the evaluation of a complex sustained STEM intervention, which drew on science capital in the intervention design and aims and on science capital research in development of evaluation methods, instruments and outcomes. The lens of science capital was invaluable within the theoretical design and set up of the STEM education programme. By considering the findings of ASPIRES about the significant predictors of future participation and identity in science, the authors were supported to identify areas of the intervention focus and were guided in making decisions about the programme's audience, reach and direction. As a consequence, the intervention chose to target younger children; work closely with a group of schools over a sustained period; engage with young people's key influencers; and highlight the transferability and utility of science through presentation of a full range of STEM careers. The theoretical model provided by science capital can thus support STEM education providers to focus interventions on these important predictors of future participation in science. This study has, however, revealed the challenges of drawing on theoretical frameworks and research instruments without due consideration of evaluation design models. This discussion thus reflects both on choices made in designing and evaluating the STEM outreach intervention and the challenges encountered in their application. Findings from the study are considered here, alongside reflections on what additional data might have brought value to answering the three research questions.

In terms of the first research question, *What are the affordances of science capital in the development of outcome measures for STEM interventions*?, the original vision of the STEM education programme was 'building science capital', using 'increased science capital' as an outcome measure for the intervention. The trajectory language of 'building science capital' has similarly been adopted by other STEM interventions applying science capital in their programme approaches (Harris et al., 2018; Nomikou et al., 2017).

The first challenge with using 'building science capital' as an outcome measure is that it had not been validated as such. It is still not known whether science capital can be built, raised, or improved by interventions and programmes, and if it can, whether and to what extent STEM education programmes can hope to counter existing societal and cultural influences to make a difference to young people's science-identities and aspirations. Greater understanding is thus required about which elements within an individual's science capital might be most malleable to intervention if future efforts are to be invested in the right areas.

Secondly, even if a young person's science capital can be built, it is not yet known what form this development might take. Creating a measure and determining a numerical science capital score assumes a linear trajectory for science capital, where a person's science capital can increase or decrease. A challenge with adopting this as a fixed perspective is that rather than placing value in the wide variety of forms of capital individuals may possess (Calabrese-Barton, 2014 in Archer et al., 2015), the implication is that some individuals possess more or better capital than others. Archer et al. now advise caution when using the terminology of 'high' and 'low' levels of science capital, in recognition that, "*important nuance is lost in translation and that the terms can unhelpfully reify and lend to unintended deficit interpretations of capital*" (ASPIRES Research, 2021, paragraph 3). Other researchers, for example Calabrese-Barton and Rahm (2014) in Archer et al. (2015) have highlighted how science capital may move horizontally as well as vertically. This raises the methodological question of whether a quantitative instrument is capable of capturing changes along a horizontal trajectory. A quantitative index, such as that used within this study, may thus be unable to represent the full gamut of possible movement over the course of a programme.

Table 5 Science capital group values for four research studies using different science capital indices

Research studies		(% of cohort)		
		High	Medium	Low
Current study (at endpoint)	England: regional survey $N=349$, 3 schools	18.4	79.3	2.3
Archer et al. (2015)	England: national survey $N=3658$, 45 schools	5.2	67.6	27.2
De Witt et al. (2016)	England: targeted survey $N=6871$, 18 schools	4.9	66.9	29.3
Christidou et al. (2021)	Norway: regional survey $N=58$, 2 schools	20.68	67.24	12.06

The final challenge encountered in applying science capital as an outcome measure, using the groupings of high, medium and low science capital, was how to interpret these findings. The science capital inspired instrument was used to generate a composite 'science capital score' which was used to categorise the young people completing the questionnaire into high, medium and low science capital groups. However, comparing the current study with two others also using a science capital index (Christidou et al., 2021; De Witt et al., 2016), showed that the proportion of young people with high, medium or low science capital varied between studies (Table 5). One reason for this discrepancy in science capital proportions could be that even though all researchers drew on science capital in development of indices and share many of the same items, the indices themselves are not the same. A second and potentially more likely reason for this discrepancy is the difference in participant sample for each study, which vary in number, location and nature of the project for participants. For example, the current study recruited schools wishing to provide STEM enrichment predominantly for young people, Christidou et al. (2021) recruited young people for a practical coding project, and the targeted survey in DeWitt et al. (2016) recruited schools participating in CPD related to science capital and comparator schools. Archer et al. (2015) reported on a national survey which although broadly representative slightly oversampled schools serving communities historically under-represented in science. This could have affected the demographic characteristics of participants involved in each project and the ensuing science capital values obtained. This suggests that funders will find it challenging to measure levels of science capital against a 'national picture' as a way of comparing the relative success of different STEM interventions. Similarly, it also suggests researchers and organisations should be cautious about citing values for standard proportions of high/medium/low science capital intended to hold across larger populations.

Thus, while the authors recognise the benefit being able to quantify and compare science capital numerically affords STEM education providers, the study raises the issue of whether using 'building science capital' as an outcome measure and attempting to quantify a broad level of science capital is consistently useful, particularly in the context of intervention evaluation. Similarly, Jones et al. (2022) who validated an instrument that includes a measure of science capital suggest that educators should use this to measure core components in STEM programmes, such as interest and career goals, and then tailor interventions aimed at building science capital using the results of this assessment.

The second research question asked: *To what extent can a quantitative index provide sufficient information to evaluate a complex STEM intervention?* The authors drew solely on quantitative aspects of science capital research in their choice of evaluation design and methods, with the intention of undertaking an outcome evaluation and creating tools to quantify the levels of science capital across different stages. Similarly, in creating the science capital index ASPIRES researchers proposed a simple, easy to administer quantitative model that could be used by the science education community in 'measuring' and determining levels of science capital at scale (De Witt et al., 2016).

Application and administration of the science capital derived research instrument proved straightforward and as the instrument recorded high levels of internal consistency and reliability, it can be considered statistically appropriate. However, analysis of data from the composite science capital scale among the cohort and baseline groups shows that since the Baseline Year 9 comparison group demonstrated statistically higher science capital than the intervention-mid, indicating that the intervention failed to produce the intended impacts. This raises the question of why? Design of the initial evaluation study did not include other forms of data that may have provided further insights into why intervention results differed from expectations.

The subsequent process evaluation undertaken using existing monitoring data aimed to understand whether failure to realise the outcomes was due to poor programme design or poor implementation. Analysis of the seven implementation factors (Humphrey et al., 2016) found several factors, where implementation was

lacking, and where intervention dose and reach were insufficient at the cohort level. It also identified great variability across the course of the programme and between schools. When used in combination with outcome evaluations, process evaluations can aid understanding of which elements of variable interventions might be having the greatest impacts on outcomes. However, to have true value they should be planned for at the outset of an evaluation so appropriate targeted data, such as interviews or observations, can be gathered for this purpose (Outhwaite et al., 2020). Overall, the study has highlighted how interpreting impacts of a complex and sustained intervention with a less structured delivery model through a quantitative index alone is likely to produce results of unknown reliability (Humphrey et al., 2016).

The third research question asks: *How appropriate was the chosen evaluation model for the evaluation of a complex sustained intervention with a flexible delivery model?* The study used a time series cohort design to consider the cumulative impact of a range of different activities over the intervention's duration, and used two age-aligned comparator cohorts drawn from the baseline to assess programme differentiation. However, several mismatches were identified between the intervention design and the chosen evaluation model.

First, the STEM outreach intervention was modelled with a 'loose enabling framework' (Humphrey et al., 2010). This flexible model was designed to explore how the intervention could work in the real-world rather than a controlled environment, this to complement schools' existing provision and strengths, and to foster the trust and strong working relationships required for sustained engagement (Education Endowment Foundation, 2019). The intervention developed a common approach at a broad level and a suitable programme of delivery activities, however, because schools could decide which intervention activities to take up, how often, when, and with which year groups, some activity types, including those that were centrally important to the ToC and considered high impact (such as teacher CPD) were not taken up as readily. The intervention design had neither identified these as critical active ingredients nor set them as part of a compulsory core offer in intervention delivery. Addressing adherence to the intended treatment model, which was a core feature of process evaluation, was thus challenged by the lack of a core treatment model to adhere to. Although evaluating the effectiveness of an intervention delivered in a loose enabling framework is possible, this may require the use of a theory-based evaluation approach or be more mixed-methods in design than is common in outcome evaluations (Outhwaite et al., 2020).

A further mismatch was that while the intervention was offered to all year groups across the school, the evaluation model tracked only one cohort, those that were in Year 7 in 2015, in Year 9 in 2017 and in Year 11 in 2019. As the actual activities delivered to this tracked intervention cohort were found to be minimal, assessing the outcomes of the intervention on a cohort who received minimal intervention proved futile.

Finally, use of comparator groups drawn from the baseline was designed to mitigate the effects of change and development within young people over the intervention's 3 years. Perhaps naively this design underestimated factors involved in implementation and the challenge of determining programme differentiation and the counterfactual. During 3 years under study, the STEM engagement programme was one small area of influence in young people's lives. Attempting to assess an intervention's impacts across a sustained timeframe, where all factors of influence on a study's population during that period can neither be described nor the variety of external influences be untangled from those generated by the intervention, proved challenging (Humphrey et al., 2016). Using this comparator model did not allow for the impact of intervention activity to be distinguished from other existing practice that took place concurrently.

Conclusions

Conceptualisation of science capital has proved invaluable to the STEM engagement community, both in building understanding of observed patterns in young people's participation in science and in guiding the thinking, direction and focus of programmes. However, choosing appropriate evaluation designs, methods and instruments remains a demanding choice for STEM engagement providers, one which first requires skills and knowledge of possible evaluation designs then secondly, the careful balancing of the rigour desired with the resources available.

Quantitative indexes, such as the science capital index, provide a relatively low cost and easy to administer instrument for STEM education interventions. However, for more reliable and insightful results, these approaches should be used alongside complementary methods, such as process evaluations and qualitative methods, such as participant or stakeholder interviews. While valuable in understanding the limitations of the designed intervention model, the process evaluation used this study could not be fully conducted due to reliance on existing process data. Inclusion of process evaluation and other complementary methods should instead be built into evaluation designs from the outset. In conclusion, and with the benefit of hindsight, the authors set out how they would proceed if including measures for science capital in future evaluation designs. It is anticipated these will be useful to other STEM educators and practitioners considering appropriate evaluation designs and methods, with the

hope that similar potential pitfalls in evaluation design and methods can be avoided.

Rather than setting broad 'building science capital' goals and providing an overall assessment or measure of science capital, the authors believe that STEM education programmes should focus on establishing more specific, targeted outcomes derived from the programme theory and use evaluation instruments specifically tailored to exploring these outcomes. A focus on specifics will better support efforts to understand which science capital elements are the most malleable to intervention. Since the outset of this STEM engagement programme, the authors have moved from the broad-brush terminology of 'building science capital', to the more focused 'supporting elements of young people's science capital'. The authors now use science capital as a guiding theory in their research, with a ToC as the working model for the intervention and to define evaluation outcomes (Davies, 2018). The science capital framework can be useful to researchers and STEM engagement professionals seeking to unpack components of science capital to consider more focused measurable outcomes (Jones et al, 2022). The measures bank developed by Engineering UK (2021) presents similar guidance to STEM education providers to consider the core outcomes of their STEM engagement, in support of survey items that are a good theoretical match for use in evaluations. Tailoring science capital in these ways will enable programme impacts to be assessed more effectively.

This study highlights the nuance involved in selecting appropriate evaluation designs and methods for sustained and complex STEM intervention designs. Adequate assessment of the effectiveness of interventions will likely depend on a combination of evaluation methods, across multiple timepoints, which then in turn will require greater investment of resources in the evaluation processes. STEM outreach practitioners can support efforts to understand what types of STEM activities and delivery models 'work', both by trialling approaches and reporting, where there is promise and also by reporting when interventions have not worked as well as, or have worked differently, from that intended. Policymakers can similarly support this, by creating an evaluation culture, where rigor is a requirement and open sharing of results is routine (McKinnon, 2022).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-023-00421-y.

Additional file 1: Analysis of individual questionnaire items among intervention and comparator groups.

Additional file 2: Principal component analysis: methodology, results and discussion.

Acknowledgements
The authors are very grateful to participating schools, teachers and young people for their sustained commitment and participation in the intervention and research study. The authors would also like to thank the anonymous reviewers for their rigorous and extremely helpful comments in the preparation of this manuscript.

Author contributions
AP and ODA were responsible for collecting and analysing data in the study. CD and RS provided oversight and guidance to the design of the research study, research methodology and interpretation of findings. The manuscript was written by AP with edits from CD and RS. All authors read and approved the final manuscript.

Funding
This work was supported by the Higher Education Funding Council of England (HEFCE) under a Catalyst Fund Grant (PD006).

Availability of data and materials
De-identified data sets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare no conflict of interest.

Author details
[1]Northumbria University, Newcastle City Campus, Ellison Building, Newcastle Upon Tyne NE1 8ST, UK.

Received: 24 March 2022 Accepted: 14 April 2023
Published online: 12 May 2023

References

APPG on Diversity and Inclusion in STEM. (2020). *Inquiry into Equity in the STEM Workforce*. https://www.britishscienceassociation.org/Handlers/Download.ashx?IDMF=d7899dce-22d5-4880-bbcf-669c0c35bda6. Accessed 28/03/2023.

Archer, L., Dawson, E., DeWitt, J., Seakins, A., & 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. https://doi.org/10.1002/tea.21227

Archer, L., DeWitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2012). Science Aspirations, Capital and Family Habitus: How families shape children's engagement and identification with science. *American Educational Research Journal, 49*(5), 881–908. https://doi.org/10.3102/0002831211433290

Archer, L., Osbourne, J., DeWitt, J., Dillon, J., Wong, B. & Willis, B. (2013). *ASPIRES: Young people's science and career aspirations, age 10–14*. https://www.geolsoc.org.uk/~/media/shared/documents/society/diversity/resources/education/young%20people%20science%20aspirations%20kcl.pdf?la=en Accessed 28/03/2023.

Archer, M., DeWitt, J., Davenport, C., Keenan, O., Coghill, L., Christodoulou, A., Durbin, S., Campbell, H., & Hou, L. (2021). Going beyond the one-off: How can STEM engagement programmes with young people have real lasting impact? *Research for All, 5*(1), 67–85. https://doi.org/10.14324/RFA.05.1.07

Askell-Williams, H., Dix, K., Lawson, M., & Slee, P. (2013). Quality of implementation of a school mental health initiative and changes over time in students' social and emotional competencies. *School Effectiveness and School Improvement, 24*(3), 357–381. https://doi.org/10.1080/09243453.2012.692697

ASPIRES Research (2021). Science vs. STEM: How does 'science capital' relate to young people's STEM aspirations? *UCL*. https://blogs.ucl.ac.uk/aspires/2021/01/15/science-vs-stem-capital/. Accessed 28/03/2023.

Banerjee, P. (2016). A longitudinal evaluation of the impact of STEM enrichment and enhancement activities in improving educational outcomes:

Research protocol. *International Journal of Educational Research, 76*, 1–11. https://doi.org/10.1016/j.ijer.2015.12.003

Biesta, G. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education, 29*(5), 491–503. https://doi.org/10.1007/s11217-010-9191-x

Black, L., & Hernandez-Martinez, P. (2016). Re-thinking science capital: The role of 'capital' and 'identity' in mediating students' engagement with mathematically demanding programmes at university. *Teaching Mathematics and Its Applications*. https://doi.org/10.1093/teamat/hrw016

Blickenstaff, J. (2005). Women and science careers: Leaky pipeline or gender filter? *Gender and Education, 17*(4), 369–386. https://doi.org/10.1080/09540250500145072

Boaz, A., Oliver, K., Cuccato, G., & Dashwood, C. (2021). *Rebuilding a Resilient Britain: Data and Evaluation Areas of Research Interest across Government*. https://www.gov.uk/government/collections/rebuilding-a-resilient-britain. Accessed 28/03/2023.

Bryan, R., Gagen, M., Bryan, W., Wilson, G., & Gagen, E. (2022). Reaching out to the hard-to-reach: Mixed methods reflections of a pilot Welsh STEM engagement project. *SN Social Sciences, 2*(2), 10. https://doi.org/10.1007/s43545-021-00311-6

Calabrese Barton, A., & Tan, E. (2010). We be burnin'! Agency, identity, and science learning. *The Journal of the Learning Sciences, 19*(2), 187–229. https://doi.org/10.1080/10508400903530044

Calabrese Barton, A., & Tan, E. (2020). Beyond equity as inclusion: A framework of "rightful presence" for guiding justice-oriented studies in teaching and learning. *Educational Researcher, 49*(6), 433–440. https://doi.org/10.3102/0013189X20927363

Carlone, H., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching, 44*(8), 1187–1218. https://doi.org/10.1002/tea.20237

Christidou, V. (2011). Interest, Attitudes and Images Related to Science: Combining Students' Voices with the Voices of School Science, Teachers, and Popular Science. *International Journal of Environmental and Science Education*, 6(2), 141–159. ERIC. https://eric.ed.gov/?id=EJ944846. Accessed 28/03/2023.

Christidou, D., Papavlasopoulou, S., & Giannakos, M. (2021). Using the lens of science capital to capture and explore children's attitudes toward science in an informal making-based space. *Information and Learning Sciences*. https://doi.org/10.1108/ILS-09-2020-0210

Claussen, S., & Osborne, J. (2013). Bourdieu's notion of cultural capital and its implications for the science curriculum. *Science Education, 97*(1), 58–79. https://doi.org/10.1002/sce.21040

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research, 60*(3), 276. https://doi.org/10.1080/00131881.2018.1493353

Craig, A. (2014). Australian interventions for women in computing: Are we evaluating? *Australasian Journal of Information Systems, 18*(2), 91–110. https://doi.org/10.3127/ajis.v18i2.849

Craig, P. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ*, 337:a1655. https://doi.org/10.1136/bmj.a1655.

Crawford, C., Dytham, S., & Naylor, R. (2017). *Improving the evaluation of outreach: Interview report*. https://pure.northampton.ac.uk/en/publications/improving-the-evaluation-of-outreach-interview-report. Accessed 28/03/2023.

Davenport, C., Dele-Ajayi, O., Emembolu, I., Morton, R., Padwick, A., Portas, A., Stonehouse, J., Strachan, R., Wake, L., Wells, G., & Woodward, J. (2020). A theory of change for improving children's perceptions, aspirations and uptake of STEM careers. *Research in Science Education, 51*(4), 997–1011. https://doi.org/10.1007/s11165-019-09909-6

Davies, R. (2018). Representing theories of change: Technical challenges with evaluation consequences. *Journal of Development Effectiveness, 10*(4), 438–461. https://doi.org/10.1080/19439342.2018.1526202

Deslandes, R., & Cloutier, R. (2002). Adolescents perception of parental involvement in schooling. *School Psychology International, 23*(2), 220–232. https://doi.org/10.1177/0143034302023002919

DeWitt, J., Archer, L., & Mau, A. (2016). Dimensions of science capital: Exploring its potential for understanding students' science participation. *International Journal of Science Education, 38*(16), 2431–2449. https://doi.org/10.1080/09500693.2016.1248520

Dromey, J. (2021). *Disconnected? Exploring the digital skills gap. Learning and Work Institute. Resource document. Learning and Work Institute*. https://learningandwork.org.uk/resources/research-and-reports/disconnected-exploring-the-digital-skills-gap/. Accessed 28/03/2023.

Education Endowment Foundation. (2019). *Putting Evidence to Work: A School's Guide to Implementation*, Education Endowment Foundation. https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/implementation. Accessed 28/03/2023.

Emembolu, I., Padwick, A., Shimwell, J., Sanderson, J., Davenport, C., & 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. https://doi.org/10.1080/09500693.2020.1729442

Engineering UK. (2021). Measures Bank. *Tomorrow's Engineers*. https://www.tomorrowsengineers.org.uk/improving-practice/resources/euk-measures-bank/. Accessed 28/03/2023.

Fixsen, D., Blase, K., Naoom, S., Van Dyke, M., and Wallace, F. (2009). *Implementation: The Missing Link between Research and Practice. NIRN implementation brief, 1*. Chapel Hill, NC: University of North Carolina at Chapel Hill. ERIC. https://files.eric.ed.gov/fulltext/ED507422.pdf.

Gardener, P. (1975). Attitudes to science: A review. *Studies in Science Education, 2*, 1–41. https://doi.org/10.1080/03057267508559818

Gokpinar, T., & Reiss, M. (2016). The role of outside-school factors in science education: A two-stage theoretical model linking Bourdieu and Sen, with a case study. *International Journal of Science Education, 38*(8), 1278–1303. https://doi.org/10.1080/09500693.2016.1188332

Harris, E., Xanthoudaki, M., & Winterbottom, M. (2018). *Tinkering and Science Capital. Ideas and Perspectives*. https://www.science-center-net.at/wp-content/uploads/2018/06/TinkeringAndScienceCapital_LR.pdf. Accessed:28/03/2023.

HEFCE, OFFA. (2014). *National Strategy for Access and Student Sources: Interim report to the Department for Business, Innovation and Skills*. HMSO. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/299689/bis-14-516-national-strategy-for-access-and-student-success.pdf. Accessed 28/03/2023.

Hoffmann, T., Glasziou, P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D., Barbour, V., Macdonald, H., Johnston, M., Lamb, S., Dixon-Woods, M., McCulloch, P., Wyatt, J., Chan, A., & Michie, S. (2014). Better Reporting of Interventions: Template for Intervention Description and Replication (TIDieR) Checklist and Guide. *BMJ, 348*, g1687. https://doi.org/10.1136/bmj.g1687

Hudson, J., Fielding, S., & Ramsay, C. (2019). Methodology and reporting characteristics of studies using interrupted time series design in healthcare. *BMC Med Res Methodology, 19*, 137. https://doi.org/10.1186/s12874-019-0777-x

Hughes, T., Nixon, I., Porter, A., Sheen, J., and Birkin, G. (2013). *Summative evaluation of the National HE STEM Programme Report* to HEFCE and HEFCW by CFE. https://www.birmingham.ac.uk/Documents/college-eps/college/stem/Summative-evaluation-national-he-stem-programme.pdf. Accessed 28/03/2023.

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook*. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf. Accessed 28/083/2023.

Humphrey, N., Lendrum, A. & Wigelsworth, M. (2010) *Social and emotional aspects of learning (SEAL) programme in secondary schools: National evaluation*. Research Report: DFE-RR049 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/181718/DFE-RR049.pdf. Accessed 28/03/2023.

Institute of Physics. (2013). *Closing Doors: Exploring gender and subject choice in schools*. Institute of Physics. https://www.iop.org/sites/default/files/2019-03/closing-doors.pdf. Accessed 28/03/2023.

Jones, M., Chesnutt, K., Ennes, M., Macher, D., & Paechter, M. (2022). Measuring science capital, science attitudes, and science experiences in elementary and middle school students. *Studies in Educational Evaluation, 74*, 101180. https://doi.org/10.1016/j.stueduc.2022.101180

Kezar, A. (2011). What is the best way to achieve broader reach of improved practices in higher education? *Innovative Higher Education, 36*(4), 235–247. https://doi.org/10.1007/s10755-011-9174-z

Long, R. (2017). *GCSE, AS and A level reform (England) Briefing Paper*, House of Commons Library. https://researchbriefings.files.parliament.uk/documents/SN06962/SN06962.pdf. Accessed 28/03/2023.
Margoluis, R., Stem, C., Salafsky, N., & Brown, M. (2009). Design alternatives for evaluating the impact of conservation projects. *New Directions for Evaluation, 122*, 85–96. https://doi.org/10.1002/ev.298
McCracken, S. (2019). *I'm a Scientist: Supporting Science Capital*. https://about.imascientist.org.uk/files/2019/11/IAS-Science-Capital-Main-Report-Sep-2019.pdf. Accessed 28/03/2023.
McKinnon, M. (2022). The absence of evidence of the effectiveness of Australian gender equity in STEM initiatives. *Australian Journal of Social Issues, 57*(1), 202–214. https://doi.org/10.1002/ajs4.142
Moote, J., Archer, L., DeWitt, J., & MacLeod, E. (2020). Science capital or STEM capital? Exploring relationships between science capital and technology, engineering, and maths aspirations and attitudes among young people aged 17/18. *Journal of Research in Science Teaching, 57*(8), 1228–1249. https://doi.org/10.1002/tea.21628
Murphy, P., & Whitelegg, E. (2006). *Girls in the Physics Classroom: A review of the Research on the Participation of Girls in Physics*. Institute of Physics. https://www.iop.org/sites/default/files/2019-04/girls-in-the-physics-classroom.pdf. Accessed 28/03/2023.
Neave, S., Wood, G., May, T., Tortis, M., Kähärä, M., Mellors-Bourne, R., Morgan, R., Desai, M., Halej, J., & Talbot, M. (2018). *State of Engineering 2018*. Engineering UK. https://www.engineeringuk.com/media/156187/state-of-engineering-report-2018.pdf. Accessed 28/03/2023.
Neumann, E., Towers, E., Gerwitz, S., & Maguire, M. (2016). *The effects of recent Key Stage 4 curriculum, assessment and accountability reforms on English secondary education*. London: National Union of Teachers and King's College London. http://downloads2.dodsmonitoring.com/downloads/Misc_Files/KingsCollege141116.pdf. Accessed 28/03/2023.
Nicolaisen, L., Ulriksen, L., & Holmegaard, H. (2023). Why science education and for whom? The contributions of science capital and Bildung. *International Journal of Science Education, Part B*. https://doi.org/10.1080/21548455.2022.2155493
Nilsen, P. (2020). Making sense of implementation theories, models, and frameworks. *Implementation Science, 3*, 53–79. https://doi.org/10.1186/s13012-015-0242-0
Nomikou, E., Archer, L., & King, H. (2017). Building 'Science Capital' in the Classroom. *School Science Review*. 98 (365), 118–124 https://kclpure.kcl.ac.uk/portal/files/70462179/Building_Science_Capital_in_the_Classroom_NOMIKOU_Accepted2017_GREEN_AAM.pdf. Accessed 28/03/2023.
Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*. https://doi.org/10.1080/0950069032000032199
Outhwaite, L. A., Gulliford, A., & Pitchford, N. J. (2020). A new methodological approach for evaluating the impact of educational intervention implementation on learning outcomes. *International Journal of Research & Method in Education, 43*(3), 225–242. https://doi.org/10.1080/1743727X.2019.1657081
Padwick, A. & Davenport, C. (2022). Lessons learned from using the National Pupil Database in the evaluation of small-scale school interventions. https://figshare.northumbria.ac.uk/articles/presentation/NUSTEM_National_Pupil_Database_Summary_Report/19294355. Accessed 22/03/2023.
Padwick, A., Dele-Ajayi, O., Davenport, C., & Strachan, R. (2016). Innovative methods for evaluating the science capital of young children. *Proceedings of the 2016 IEEE Frontiers in Education Conference*. 1–5. IEEE. https://doi.org/10.1109/FIE.2016.7757680
Peterson, A. (2016). Getting 'What Works' Working: Building blocks for the integration of experimental and improvement science. *International Journal of Research & Method in Education, 39*(3), 299–313. https://doi.org/10.1080/1743727X.2016.1170114
Powell, A., Neilsen, N., Butler, M., Buxton, C., Johnson, O., Ketterlin-Geller, L., Stiles, J., & McCulloch, C. (2018). *The Use of Theory in Research on Broadening Participation in PreK–12 STEM Education*. Community for Advancing Discovery Research in Education (CADRE). https://www.edc.org/use-theory-research-broadening-participation-prek%E2%80%9312-stem-education. Accessed 28/03/2023.
Reed-Rhoads, T. (2011). Assessing K-12 Outreach. *MRS Bulletin, 36*, 264–269. https://doi.org/10.1557/mrs.2011.62
Reinholz, D., White, I., & Andrews, T. (2021). Change theory in STEM higher education: A systematic review. *International Journal of STEM Education, 8*, 37. https://doi.org/10.1186/s40594-021-00291-2
Reynolds, J., Di Liberto, D., Mangham-Jefferies, L., Ansah, E., Lal, S., Mbakilwa, H., & Chandler, C. I. (2014). The practice of 'doing' evaluation: Lessons learned from nine complex intervention trials in action. *Implementation Science, 9*(1), 1–12. https://doi.org/10.1186/1748-5908-9-75
Rosicka, C. (2016). *Translating STEM education research into practice*. ACER. https://research.acer.edu.au/professional_dev/10/. Accessed 28/03/2023.
Rossi, P., Freeman, H., & Lipsey, M. (1999). *Evaluation: A systematic approach* (6th ed.). Sage.
Sadler, K., Eilam, E., Bigger, S., & Barry, F. (2018). University-led STEM outreach programs: Purposes, impacts, stakeholder needs and institutional support at nine Australian universities. *Studies in Higher Education, 43*(3), 586–599. https://doi.org/10.1080/03075079.2016.1185775
Sarmiento-Márquez, E. M., Pishtari, G., Prieto, L. P., & Poom-Valickis, K. (2023). The evaluation of school-university partnerships that improve teaching and learning practices: A systematic review. *Educational Research Review*. https://doi.org/10.1016/j.edurev.2023.100509
St. Clair, T., Hallberg, K., & Cook, T. D. (2014). *Causal Inference and the Comparative Interrupted Time Series Design: Findings from Within-Study Comparisons*. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED562724. Accessed 10/03/2023.
Sullivan, G. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education, 3*(3), 285–289. https://doi.org/10.4300/JGME-D-11-00147.1
Wells, M., Williams, B., Treweek, S., Coyle, J., & Taylor, J. (2012). Intervention description is not enough: Evidence from an in-depth multiple case study on the untold role and impact of context in randomised controlled trials of seven complex interventions. *Trials, 13*, 95. https://doi.org/10.1186/1745-6215-13-95
Wilkerson, S. B., & Haden, C. (2014). Effective Practices for Evaluating STEM Out-of-School Time Programs. *Afterschool matters*, 19, 10–19. ERIC. https://files.eric.ed.gov/fulltext/EJ1021960.pdf. Accessed 28/03/2023.
Yang, Y., & Land, K. (2013). *Age-Period-Cohort Analysis: New models, methods, and empirical applications*. Taylor Francis Group.
Ziegler, R., Hedder, I., & Fischer, L. (2021). Evaluation of science communication: Current practices, challenges, and future implications. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2021.669744

Publisher's Note