Policy Support Fund (PSF) 2024-25 funded project: Feasibility of enhancements to the Office for Students (OfS) Standards of Evaluation Evidence for Access and Participation Plans

REPORT 1

RESEARCH REPORT & RECOMMENDATIONS TO THE OFS

AUGUST 2025

Lead Author: Joanne Moore
Email: jm3196@bath.ac.uk

Annette Hayton
Email: arh39@bath.ac.uk

Institutional Partners:
Leeds Conservatoire
London School of Economics and Political Science
Loughborough University
University of East London
University of Hertfordshire
University of Law
University of Sussex

Project Reference Group:
Anna Anthony, Director of HEAT
Matt Dixon, Professor of Economic & Social Policy, University of Bath
Manual Madriaga, Professor in Education, University of Nottingham
Tony Moss, Pro Vice-Chancellor Education and Student Experience, London South Bank University
Liz Thomas, Research Centre Lead: Centre for Research on Education and Social Justice (CRESJ), University of York
Kathleen M Quinlan, Professor of HE and Director of Centre for the Study of HE, University of Kent

CONTENTS

1 INTRODUCTION

This report sets out findings of the Policy Support Fund (PSF) 2024-25 funded project: Feasibility of enhancements to the Office for Students (OfS) Standards of Evaluation Evidence for Access and Participation Plans. The OfS standards of evidence[1] were developed to support evaluation capability building in higher education and to support decision-makers in making consistent judgements when assessing evidence about the effectiveness of a particular policy, practice or programme.

Effective evaluation is important to identify which interventions are making a difference, and those which are not having the desired effect, in order to ensure that access and participation monies are used in the most effective ways that benefit outcomes. Understanding the impact that access and participation interventions are having matters if we are to identify effective practices and contribute to knowledge about what works to deliver improved widening participation outcomes and impacts for students in higher education.

The aims of the research were to:

- Support universities' capacity to produce effective evaluation evidence for widening participation interventions and fulfil their Access and Participation Plan (APP) commitments.
- Develop recommendations and tools enabling the transfer of effective practices across the higher education sector.
- Inform national OfS guidance on evaluation methods supporting transfer of best practice from impactful programmes.

This report should be read in conjunction with *Report Two: Practices for Evaluation Strengthening: Learning from the research partners*, which discusses the challenges for institutions of APP evaluations, highlights areas of development to strengthen the evaluation approaches, in context, provides examples from the case institutions of planned evaluations and describes their evaluation capability strengthening activities.

1.1 Background

The evidence base for access and participation interventions has been criticised for being underdeveloped nationally, which has raised concerns about the extent to which what is being delivered is based on evidence of impact.[2] One review concluded that widening participation tends to be activity-led - rather than outcome/theory led - and concerned with data generation rather than critical thinking (Austen et al. 2021).[3]

Standards of evidence are not new as a tool for improving practice and effectiveness by learning from publicly funded initiatives. Numerous frameworks have been developed to help structure how evidence is collected, interpreted and assessed.[4] The standards of evidence in higher education were first published in 2017 as a result of a joint HEFCE/OFFA initiative, with a focus on evaluation of outreach interventions, particularly as a strategy to raise aspirations and attainment of young people from groups under-represented in higher education (Crawford et al., 2017).[5] In 2019, the OfS adopted and extended the standards as part of their general access and participation plan guidance[6]. The standards promote transparency and accountability by providing a shared reference framework. They support evidence-based decisions about which interventions are effective in generating desired outcomes and impacts, and therefore the best use of access and participation resources. They are based on three types which generate different kinds of evidence of impact:

- Type 1) narrative evaluation - knowing what will generate impact and why (including existing evidence of effectiveness);

[1] https://www.officeforstudents.org.uk/publications/standards-of-evidence-and-evaluating-impact-of-outreach/

[2] Blake, J. (2022, 8 February) Next steps in access and participation. Speech given by John Blake, the Office for Students' Director for Fair Access and Participation. https://www.officeforstudents.org.uk/news-blog-and-events/press-and-media/next-steps-in-access-and-participation/

[3] Austen, L., Hodgson, R., Heaton, C., Pickering, N., Dickinson, J., Mitchell, R. and O'Connor, S. (2021) Access, retention, attainment and progression: an integrative review of demonstrable impact on student outcomes. York: Advance HE. https://www.advance-he.ac.uk/knowledge-hub/access-retention-attainment-and-progression-review-literature-2016-2021

[4] Some examples include the GLA's Project Oracle (for youth provision); NESTA's standards (for innovation funding); and Reclaiming Futures (for justice system reform).

[5] https://pure.northampton.ac.uk/ws/portalfiles/portal/6246443/Crawford_Claire_UoN_2017_The_Evaluation_of_the_Impact_of_Outreach.pdf

[6] https://www.officeforstudents.org.uk/publications/standards-of-evidence-and-evaluating-impact-of-outreach/

Type 2) empirical enquiry - evaluation to measure the difference made by activities and practices compared to what might otherwise have been expected to happen;

Type 3) causal claims – to identify whether the outcome and impact was a direct result of the activities.

Experience shows that using evidence to improve practice and decision-making is much more likely to happen when the environment for change is right. The OfS evaluation self-assessment tool[7] allows providers to assess the conditions in place internally for impact evaluation and to identify steps for improvement in relation to four dimensions of their evaluation work: the strategic context; programme design aspects; evaluation design aspects; and the framework for implementing and learning from evaluation.

1.2 Why was this project undertaken?

The period since the original standards of evidence were first developed has been a time of significant change in the English higher education landscape. The establishment of the OfS, has brought together regulatory and funding levers. Access and Participation Plans (APPs) operate alongside the general 'conditions of registration' (minimum expected performance measures) which funded providers must conform to (OfS, 2022)[8], and the monitoring of equality, access and participation, and quality assurance functions. Teaching excellence is a central theme in the accountability discourse, and an important aspect of quality assurance. The overlaps between APP and other internal QA and external (regulatory) mechanisms have become more obvious. Coupled with this is increasing concern for student involvement and the importance of demonstrating a 'whole provider approach' (WPA). The OfS continues to emphasise external (as well as internal) knowledge development with requirements to publish evaluation outputs. The Higher Education Evaluation Library (HEEL) is being developed as a repository for sharing evidence.

In this context, the research sought to identify the effect of standards of evidence on current approaches to evaluation, consider how evidence standards for impact evaluation can be extended and enhanced to take account of contextual factors, and to identify evaluation approaches that have most influence in different institutional contexts, in order to produce recommendations, and materials to support learning from evaluation of widening participation interventions.

1.3 How was the research undertaken?

This was a collaborative project involving in-depth research and a collective consultation process with seven case institutions which were chosen to represent the diversity of provision across the English higher education sector. A reference group supported the reporting and was involved in agreeing the outputs. Information on the research partners and methods are given in Annex 3.

1.4 This report

Section 1 recaps the background, aims and objectives of the project and the approach taken to undertaking the research.

Section 2 draws out findings from the research in relation to how providers have engaged with standards of evidence, the issues emerging for evaluation capability development.

Section 3 discusses how the standards are conceptualised in the field and the types/levels discourse underpinning the application of evidence standards.

Section 4 explores developments in the context for access and participation, the provider level considerations, and resulting implications for the application of the standards for evidence-based practice.

Section 5 provides some conclusions and sets out recommendations emerging from the project for the OfS.

Annex 1 describes the research method.

Annex 2 suggests potential enhancements to the standards framework.

Annex 3 suggests an approach to integrating impact evidence with other types of evaluation.

[7] https://www.officeforstudents.org.uk/for-providers/equality-of-opportunity/evaluation/standards-of-evidence-and-evaluation-self-assessment-tool/

[8] OfS (2022d) Conditions of registration. https://www.officeforstudents.org.uk/advice-and-guidance/regulation/registration-with-the-ofs-a-guide/conditions-of-registration/

2 IMPLICATIONS OF THE STANDARDS OF EVIDENCE

This section discusses ways in which the standards have influenced access and participation evaluation work. The findings are for providers included in the research, however, there was significant commonality in themes emerging, and given the diverse sample, this suggests these insights are likely to resonate across the English higher education sector as a whole.

Key Findings

The research highlighted that the standards of evidence are supporting communication regarding the evaluations taking place within higher education providers to review access and participation work, and that they have played a role in evaluation capability building. The standards have underpinned developments in terms of the overall evaluation strategy; in relation to decisions about how to evaluate specific interventions are made and the types of evaluation that are put in place; and in terms of developing understanding of what counts as useful evidence for decision-making amongst practitioners and evaluators.

The standards of evidence and the associated self-assessment tool have pushed providers to implement increasingly robust approaches to assessing what difference their interventions are making to outcomes and getting greater surety that what they are delivering is making a difference. The standards support the decision making processes at project, programme and institutional levels. They support planning and have contributed to good practice approaches particularly in relation to building the rationale for interventions and obtaining clarity of outcomes and impacts using theory of change.

The standards provide a '*common language*' and a '*touchstone*' for the sector on evaluation. However, there is also the danger that evidence types tend to be conceptualised rather broadly, which may limit the extent to which the standards can be practically useful for the purpose of further impact evidence strengthening across the sector. There is a danger of the standards becoming diluted and a 'shorthand' for a methodological distinction between experimental, quantitative and qualitative methods and are applied not just to impact evaluation but evaluation in general.

2.1 Feedback from the key informants on the implications of the standards of evidence

2.1.1 A 'common language' when referring to evaluations

The standards of evidence were described as a 'common language' and a 'touchstone' for the HE sector. Overall, there was a strong sense that colleagues recognise the importance of standards and are making efforts to strive for increasingly stronger evidence of the difference that investment in access and participation is making.

> *The sector as a whole I'd say has moved from student quotes - about the lovely time etc – to prove more and more that we are doing things for good impact. It forces us to the rationale, and to focus on impact as well as outcomes.*
>
> *I think there is that broader understanding of different types of evaluation and how in building our evaluation capacity and 'journey' to evidence causality we move from everything being narrative. I think there's that broader understanding there.*

The standards have brought to the forefront of thinking the use of evidence to assess the difference being made to student and other outcomes and how to use the evidence to make decisions on APP interventions using evaluation evidence.

> *It helps the* [Widening Participation] *Committee know we're producing things to the way we should be producing them. That we will take next steps and at the same time look at the 'so what' of all of this: what it means for us. And will make sure we're using evaluation and not just doing it for the sake of it.*
>
> *It has helped us feel secure and be more sophisticated with what* [evaluation] *we do.*

The standards have been used to underpin discussion about how different approaches to evaluation are possible, with an emphasis on making sure that consideration is being given to capturing the outcomes generated by an intervention and the difference the intervention has made. However, beyond this general concern, it is clear that in-common ideas of what the evaluation types denote is conceptualised in a rather general way. Different audiences engage with the language of the standards to a greater and lesser extent.

> *You don't necessarily refer back to it as a document to be honest, but we sort of think broadly in terms of an evaluation type, and because TASO align perhaps broadly with that evaluation types, we've drawn on them. We use the knowledge from them rather than as a document that we refer back to now in practise.*
>
> *I'm not sure if you said to someone explain to me a Type 2 or give me an example of a Type 2 evaluation, whether they'd be able to articulate it, but I think that there is that broader understanding of different types of evaluation*
>
> *That language is not as common in our institution internally because that's probably to do with evaluation resource and capacity and it might only be the people who are very involved in access and participation evaluation who really understand. If you said, oh, it's a Type 2 or a Type 3, it's only something that's broadly understood, even within our data, research and evaluation working group.*

Practitioners and policy makers in general are interested in finding out why as well as what is making a difference, so the interest goes beyond proving causality per se. In practice the standards of evidence encompass a range of different evaluation purposes, not impact evaluation, because conceptions of Type 1/2 evidence – particularly insights about practice - support the wider narrative about why an intervention might be considered to be successful (e.g. drawing on different types of implementation and process evaluation).

> *I think there needs to be some sort of recognition that [the standards] are looking at things from an impact evaluation perspective and not necessarily all the other things that evaluation is going to do for you and how you're going to use it* [evidence].

2.2.3 A driver for impact evaluation capability development internally

The research identified that standards of evidence had been important within all types of higher education providers in framing conversations about evaluation, with the clear conclusion that the standards have helped to push providers towards more rigorous and considered evaluation of outcomes and impact (and certainly away from relying solely on process-related indicators and measures). The standards are not the only aspect of the OfS's evaluation capability building endeavour, but they are a prominent feature and key tenet in what providers are being asked to do. Comments at interview attested to active engagement with the standards as a tool for making improvements and the concern to reach the best possible solution for APP evaluation to inform institutional decision-making.

> *It's the 'so what?' that's important: we've assessed ourselves but now what? At a scrutiny level that involves change – culture change – and having that touchstone makes front and centre that need for causal evidence. We would have done what we did and eventually would have had that reflection – but the* [standards] *framework forces us to get there quicker.*

The standards appear to have been instrumental in encouraging more consistency in the approach to collecting evidence within providers, and in the systems to inform the institutional understanding of the access and participation work going on.

> *We went from 'who knows?' to proper tracking, and competent people to ensure everyone's doing things in the same way.*

In driving forward the overall evaluation effort in provider organisations, the research suggests the standards put the focuses on developing expertise and capacity. There is a trend towards professionalisation of evaluation roles and an emergence of dedicated institutional evaluation posts (evidenced for example by the increase in job adverts for evaluation posts). There was a sense from fieldwork that dedicated posts and specialised expertise were a pre-requisite to meeting heightened expectations for evaluation of access and participation activities.

> *Professionalisation is absolutely necessary – it's not fair to lump that on a grade 4 widening participation manager* [re expertise in quantitative data collection].
>
> *It's only really the last few years that we've got someone who really understands evaluation and how to do it properly*.

A common theme across all cases was that widening participation evaluation is playing a major role in driving forward the overall evaluation effort in the case providers and is supporting the development of a wider institutional culture of evaluation. APPs have become an institutional catalyst for evaluative change, including for 'on course' interventions. This shift has facilitated wider planning/thinking around alignment with institutional priorities. APP evaluation seems increasingly important for broader institutional mechanisms for reporting and decision-making (i.e. in addition to being crucial for decision-making for APP specific activities). Key informants spoke about overlaps between APP evaluation and a range of other regulatory mechanisms (TEF and B3 conditions), and linkages to quality assurance processes. Examples included – work in student services; mental health and well-being initiatives; strategy creation/co-creation activities with students; curriculum development initiatives (amongst other things). Overall there was a sense that the APP evaluation effort was driving forward the whole institutional approach to evaluation internally.

> *It's shifted the language subtly.* [In relation to APP evaluation] *we try to say 'you may not think this is applicable, but it is good practice'. It doesn't mean it's going to work for all. This approach can capture interest. We have to make it* [evaluation] *a university-wide applicable thing. It helps the university's strategic plan. We have to do it for TEF. They're the same.*
>
> *The TEF is interested in what's highly effective. So straightaway there's one type of outcome which is about effectiveness. But throughout* [evaluation is needed]*, it's not sufficient to just say what we are doing.*
>
> *We want our students to be successful and if measured by B3 to get to the second year, then to get a good degree and a good graduate outcome. To get that right we need to triangulate and pull in evidence holistically.*
>
> *We build the evaluation on-top of the existing processes and data collection and co-ordinate and facilitate school (i.e. department) level priorities. Big projects fit across different teams and learn from each other.* [The evaluation work] *is not isolated.*

As well as links between APP evaluation and internal quality assurance processes such as service level and curriculum reviews, there was links to academic practices, especially reflexive practices.

> *The big piece of work which is ongoing is brokering evaluation into existing services. When designing it rather than inventing something new, we want to use the evidence to see how effective what we already do is, to collect more data to remind us why we're doing it and make it better. We need to establish a baseline before we reinvigorate and will be aiming for a mix of Type 1 and Type 2, neither of which are historically well established.*

At the same time, some key informants were cognisant of the existence currently of a highly dynamic and fluid context where the sustainability of evaluation developments could not be taken for granted. Ultimately decisions affecting evaluation implementation plans could be outside the control of the people responsible for the evidence generation.

> *We've got in-house expertise but to drive strategically where we need to be and create a culture of evaluation will take more than strong will and pockets of good practice. We've not nailed the whole institutional approach to evaluation – what the landscape will look like for the next four years is unclear. We've published our intent but how we deliver it is ever changing.*

2.2.4 Supporting evaluation decision making

The standards were playing a role in decisions on intervention-specific evaluations and their implementation. This was often played out in the conversations which evaluators and colleagues with accountability roles were having with practitioners and other service delivery staff who are usually the ones with responsibility for embedding an evaluation.

> *It's helpful. They really do set a basic level of expectation which we draw upon when speaking about what the standards are. It's not only a case of Type 3 – the more of that from OfS the better – but tools at the start of the conversation to support the work – what the baseline is in practice.*
>
> *Ultimately there are strengths/weaknesses and we thought about this. That helped us reflect on what we're designing and gives consistency in how we chart/look at standards – what the best approach should be for an intervention OR whether the test or tool being used* [to evaluate] *is right? Great to have something that automates decisions on it.*

In one case the standards were being used as part of an assessment regarding the current state of play on evaluation (i.e. when it comes to assessing what is already happening in relation to existing and ongoing initiatives). This example highlights the (not necessarily uncommon) situation where some (devolved) activities within the remit of APP needed to be brought within the remit of teams leading on APP accountability. In this case work was going on to the ensure APP activities were coming in scope of agreed access and participation evaluation procedures for the forthcoming APP round.

> *Useful for us to monitor at programme level where it is in terms of evidence being created. However, it's not an evaluation until it's got an evaluation design on it.*

2.2.5 Helping to prioritise evaluation work

The standards offer a framework for considering what type of evaluation is proportionate for different kinds of intervention, and direct attention to the most resource-intensive programmes such as long-term or multi-activity interventions which require stronger evidence of impact than 'light-touch' ones because it would be more risky to continue to devote the level of resource unless the activity can be shown to be having the beneficial impact it is aiming for. As might be expected, considerations such as the availability of existing evidence of impact also affects the approach in terms of meeting the information needs of the intended evaluation users.

> *Standards are helpful for internal conversations. A shared language for getting internal stakeholders who are doing evaluation to think about where and how. Where are different types needed and how much is needed in different places?*

Part of the challenge for APP staff in accountability roles (i.e. with responsibilities for reporting internally and externally) is keeping track of all the interventions going on, ensuring ongoing monitoring and accountability reporting, ensuring those delivering interventions are supported to evaluate and ensuring that appropriate resource is going into strengthening the impact evidence. Getting clear on the priorities for evidence gathering has helped to organise the work.

> *The APP brought all these random activities into one focus. Standards are used as a way of making sense of different levels of evidence and the difference between good and bad. Brings order and purpose, direction, focus and control.*

> *We're more focused now: we know when and how each activity will be evaluated and when and how we'll get the results.*

Key informants in evaluation roles would like to see more support for using evaluation resource to the best effect overall as part of access and participation planning. One key informant commented on some ambiguity in OfS guidance:

> *Regulatory Notice 1 says all APPs have to have credible Intervention Strategies which explain how each of the outcomes will be evaluated - for us that's 90-100 interventions. Regulatory Advice 6 says we should be prioritising evaluation according to what we already know and what we do not know about.*

Uncertainty existed around expectation to evaluate an intervention at the project/programme level or by setting in place an overarching approach for the entire intervention strategy (or both). The comments highlight the tension for evaluation planning on the ground: on the one hand there is a clear need to prioritise (limited) evaluation resources to be most effective and to make sure that the evidence can be used to the greatest effect; on the other hand, there is a need to strive for strong evidence across the field (at least Type 2 for intervention strategies as a whole).

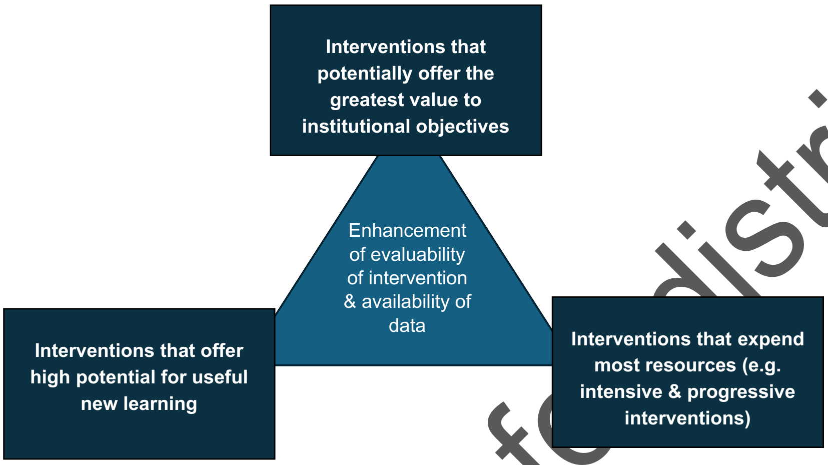
> *What we wanted to do with the existing plan is to get the culture embedded – dedicate time, send out toolkit resources or run network events - but there's only so many conversations you can have.*

Where the Intervention Strategies included many diverse strands of work it makes sense to put the efforts into enhancing the evaluability of the interventions, for example through identification of indicators to capture the outcomes and impacts and what data would be used to measure these, as part of a proportionate approach. At this stage in the APP planning process, much of this work was being pushed forward through working with practitioners to agree logical frameworks for interventions and/or an enhanced theory of change. Decisions on which interventions to focus on could take in a range of considerations (see Figure 2.1). Most effort to enhance evaluation (for example by testing the evaluability and securing access to data) was being direct at resource-intensive interventions, linked to overall objectives, and offer innovation and new knowledge.

> *Anything that gets a lot of resource, the person in charge has the loudest voice, because that's pitched at high level evaluation – because it's an important activity. Some activities are smaller and they're quite quiet and that's OK.*

> *On legacy funding schemes there's the choice to whittle down something or articulate resources into APP goals or find activities that are worthwhile evaluating.*

Figure 2.1: Criteria used to focus efforts to support stronger evaluation designs


Interventions that potentially offer the greatest value to institutional objectives
Enhancement of evaluability of intervention & availability of data
Interventions that offer high potential for useful new learning
Interventions that expend most resources (e.g. intensive & progressive interventions)

Some tension has emerged because providers are being asked to make at least a minimum APP commitment to Type 2 evaluation across the board as part of their Intervention Strategies. In practice it could be hard to evaluate all activities to the same extent all the time (although successively strong evidence could emerge over the lifetime of the APP); for example, for remedial activities where there is no obvious comparison group. Light touch activities might not warrant the level of data collection needed for a Type 2 evaluation. Another key issue is whether some interventions are designed to have an indirect benefit – so the impact is less direct in terms of being able to definitely identify the cause and effect outcomes and impacts.

> *Some interventions – such as sessions (for staff) on gender equality - it's not in our control to do it* [evaluation]*, that's a capacity issue.*

Not every activity needs to be evaluated to the same extent every year – and for some interventions evidence of good impact might already be established in the existing evidence base or the available institutional level evaluations (Type 1), especially if it was delivering only a 'light-touch' contribution to the overall objectives. The principle of proportionality in the standards of evidence supports the view that low level and light tough interventions might not warrant the resource involved in Type 2/3 evidence generation. At the same time, the need for Type 2 evidence implies evaluating how interventions work together to deliver a beneficial outcome.

3.2.5 Supporting good practice approaches to impact evaluation

Alongside other evaluation guidance and tools, the standards support good practice approaches to intervention planning – such as having an evidence based for what is delivered and being clear on the expected results. An expectation for Type 1 evaluation has led providers to examine the evidence base to ensure a rationale for the interventions taking place. Building the impact narrative was at the forefront of thinking, although the extent to which providers drew on external evidence varied.

> *There's not time for full blown literature reviews, and anyway there's only limited evidence. Our strengths come in the work we do with the students we work with – which points back to the same thing – why are we doing that* [intervention]*?*

Developments in the use of theory of change as an intervention and evaluation planning tool is helping providers to achieve greater focus and clarity on the hypothesises/expected causal-linkages underpinning interventions and associated outcome and impact measures. All the cases were using theory of change in some way, putting attention on the need to specify outcomes and impacts precisely and understand how these will be achieved. Evaluators were increasingly involved in intervention design and planning for new and existing initiatives, which is contributing to evaluation capability building.

> *You can't evaluate if you don't know what's being delivered – that's why TOC's important.*

> *Using a logic model enables us to attribute individual impact to a programme (context, assumptions and mechanisms). This is a change in culture. Before it was: no context, no strategy, 1-2 outcome measures, no benchmarking, no assumptions. So it's been quite key to use TOC – this is what we're aiming for and why.*

> *We didn't make funding decisions but need a good idea of why interventions were commissioned – we ask the questions about the rationale and what it'll achieve, we get answers out of project leads.*

A benefit of TOC currently, given the stage of APP implementation, was in terms of the rationale for interventions and the focus on gaining clarity on outcomes and impacts.

> *Getting a TOC in place and having basic monitoring data – that's a good success at this stage.*

The main value is in the process of reflection and deliberation on the activities (rather than the TOC as an end in itself). TOCs were helping to generate Type 1 evidence whether this was existing evidence, learning from practice, and/or drawing on practitioner insights through TOC building sessions.

> *My role is important for the design of TOCs – I do a lot of admin for it – and have capacity to do research. This is part of evidence base-building – literature reviews – making sure there's an evidence base behind it. That for everything there's a reason, policy or approach rooted in research. This lends to creating TOCs more easily – knowing where we're going and what we'll get out of it.*

> *The TOC helps to see different outcomes for different stakeholders, which we can break down in the evaluation. Most importantly, what are the student outcomes when they're involved in design – what do they want to get out of it?.*

Use of theory of change work was clearly boosting the Type 1 impact evidence available to be considered in formative intervention decision-making. It is also likely to have implications for theory-based evaluation to summatively assess impact – although this will probably evolve over time and is not yet reflected in the standard of evidence types.

> *I'm excited about TOC – it catches the thinking. You can show a more considered way of thinking about things and what could have more impact, rather than just how good you've been and justifying the money you're already receiving.*

The feedback from key informants suggested that an upcoming challenge will be making sure the theory of change models would be 'living documents' and a decisive part of capturing the learning which would enable ongoing evaluation (i.e. to inform theory-based evaluation approaches). In the short term having clarity on outcomes is expected to help with the collation of outcome evidence – and how to use theory of change for testing impact generally needed further work.

> *I think saying it's a living document and it being a living document are not necessarily the same thing. At the moment I think the primary purpose for us is to get our thoughts in one place about what we're doing. Why are we doing it? What do we hope to achieve? That's all informed how we measure the outcomes. We'll review the outcomes, then we'll report back. Then it kind of gets a little bit lost in the big decision about whether we want something for another year or whatever. That more granular and nuanced sort of learning from an intervention that's really beneficial is less formal and sometimes gets a little bit lost.*

The use of theory of change for theory-based evaluations is starting to be developed for APP interventions. The quantity and quality of TOCs being developed seems impressive and this work has already developed some valuable new insights into the mechanisms underpinning education outcomes, and the difference that particular practices can make. Going forward there needs to be further discussion of the implications for the evaluation approach – should evaluators take a 'black box' approach to assessing the outcomes and impact contained in the TOC; or should they aim for more theory-based evaluation methods which hypothesise about the programme theory as well the outcomes? How to use TOCs in the planning-evaluation-planning cycle over time was an issue where key informants felt there could be more help/guidance.

> *Maybe there's a gap there in the guidance about how practically you do that* [evaluate the programme theory]. *Maybe a little bit more support with 'how?'. What that iterative evaluation cycle looks like in practise and how to ensure that theory of change are living documents.*

Case providers were putting efforts into good practice developments in other areas of practice – such as follow-up and tracking. A prominent example was ensuring to attach participant data to engagement in interventions – i.e. linking participation to outcomes and to opportunities for tracking of participants (e.g. via HEAT) (recognising the benefit of having participant tracking as a potential means of opening up opportunities for counterfactual evidence and quasi-experimental designs).

2.2.6 Supporting individuals/teams on evaluation

Several respondents commented that the standards had been a useful starting point when coming into an institutional evaluation role (and perhaps especially for those coming from a practitioner route with limited prior evaluation expertise to draw on).

> *I think they are useful, personally. They were more useful when I was new to evaluation, but they came at the same time that I started, as an evaluator. They were very, very useful at that point. It's not something I really refer people to now, I think there's perhaps more recent guidance.*
>
> *I do think standards of evidence are useful. The problem in an institution like ours is we don't have lots of people who just do research or can take time off to do it. Guidelines and frameworks are important to us, otherwise we've got to reinvent the wheel and start from scratch.*

Colleagues in evaluation roles were supporting practitioners to build understanding and skills in evaluation. The standards guidance could perhaps be more helping in supporting development of expertise across teams. One respondent in a dedicated evaluation role commented:

> *I've got the table of the standards of evidence. I've built a SharePoint page for evaluation which is available to all staff at the university, so I've got that initial kind of table, with the evaluation types. But I think that's something that we could build on more if the standards are updated and more relevant to the current context and expectations and that kind of thing.*

Another key informant said that communication of standards to practitioners was hampered by the existence of a hierarchical view of evaluation designs – because more robust designs did not reflect the way the institution was evaluating (discussed further below); plus in their current format the standards were not considered very user friendly or comprehensible to practitioners.

The scale of the task of building evaluation culture and skills at the coalface of access and participation delivery was a common theme in the interviews.

> *We've got a handle on own* [widening participation] *activity as it fits in our team. Making that institutional – the learner progression framework and whole institutional stuff – that's a step change we're keen to make but we're not going to evaluate everything. That's impossible with one person. The expertise isn't there, just some pockets.*

Having dedicated evaluation posts is helping APP colleagues to make judgements about the utility and effectiveness of different evaluation designs in meeting evaluation purposes in different contexts. At the same time, the bulk of the evaluator job description is focused on supporting those at the delivery-interface to plan and undertake evaluation work, with only a minority of their time available for the business involved in actually undertaking evaluations. There were calls for guidance to support capability building at practitioner level, especially guidance to support decision making on applying the standards in practice. A key informant in a lead evaluation role commented in relation to supporting practitioner to embed evaluation as part of their activities:

> *We need standards and frameworks, so long as they help in a practical way and not put people off.*

Guidance on how to engage with different methods which support evaluation strengthening and higher quality data and evidence would be particularly beneficial to evaluators and practitioners with limited prior evaluation expertise. The sense emerged that it's important that practitioners retain skills and expertise in evaluation - across all kinds of methods - and there was a danger that the 'stronger' types could potentially be seen as more 'academic' and only in the remit of experts. A practitioner key information commented:

> *We need more in the way of being able to see how different types of evaluation may be practically delivered.*

Standardised tools including validated questionnaires which support evaluation of student outcomes are considered to be desirable not only to prevent evaluators having to 'reinvent the wheel' but because they help to build confidence in evaluation amongst practitioners delivering interventions and can promote deeper reflection of the conceptualisation of different types of psycho-social outcomes and the mechanisms underpinning achievement of these.

Sector level tools and methods which bring stronger types of evidence into the hands of practitioners were welcomed (for example, the HEAT comparator tool and progression to higher education reports). Sector resources which give concrete methodological guidance and examples were also well received (e.g. the 'small-n' guidance developed by TASO[9]).

2.2.7 External communication

Key informants said the standards were useful in communication externally to the OfS and sector colleagues about the evaluation plans and the types of evidence being collected. The comments suggest that the standards are welcomes because there is a reassurance benefit from having a framework to categorise the expectations for outcome/impact evaluations which can be justified within the organisational context.

For sector discussions and sharing the standards underpinned conversations about evaluation methods for Type 2/3 evaluations. However, when it comes to delivering Type 3 evidence some key informants were conscious of the methodological and practical difficulties and would only be looking to do this in exceptional cases.

> *I'd say we're comfortably into the middle zone, right now. Type 3 is not realistic so we're staying at Type 2 for now. For the time being we lack capacity in data and evaluation.*
>
> *We sit in the correlation space to see what works really without the need to push for Type 3.*

Having to set the expectations for impact evidence generation across each strand of activity in the APPs was not a requirement that was especially welcomed. There were various tensions including pressure to commit to certain types of evidence in advance (i.e. at least Type 2) before the requirements for different types of evaluation (e.g. access to data) were fully understood. Key informants said the type of evaluation needs to be linked to the specific type of outcomes and questions for the evaluation in each case; that the type of evaluation evidence could change over time; that the evidence once collected might not meet the standard agreed despite the best intentions (e.g. an inconclusive result on an RCT); and that the data sources that underpinned the approach might not be fully guaranteed (e.g. schools data or access to student records data). Another potential issue is that qualitative outcomes and impacts might be harder to quantify in a way that would enable a useful counterfactual to be built to meet the criteria for a Type 2 evaluation. Plus the longer term quantitative outcomes might not be identifiable in the data due to data issues and confounding factors. These issues are playing into some sense of uneasiness because the APPs have not started yet. However, there was very much a feeling that best endeavours were being made to achieve the most methodologically sound evaluation as possible in context.

2.2 Use of the standards of evaluation self-assessment tool

The evaluation self-assessment tool allows providers to self-assess on the application of standards of evidence within the context of the systems and structures in place internally for evaluation (strategy, planning implementation and learning).

Overall, colleagues seemed to be in favour of having a tool to help them reflect and review the context for evaluations and evidence generation:

> *The self-assessment surfaced where our gaps are. Ultimately, there are strengths/weaknesses and we thought about this: that helped us reflect on what we're designing and gives consistency in how we look at standards.*

[9] https://taso.org.uk/wp-content/uploads/2022-06-17_Impact-evaluation-with-small-cohorts_methodology-guidance_TASO.pdf

In one case institution, working towards improvement in the self-assessment score had been made a stated aim and part of the strategy for evaluation capacity/capability development internally. There had been leverage for additional dedicated evaluation staff roles as a result here. In another case a new evaluation staff role had been created following a review of evaluation capability which had referenced the standards.

Overall, there was mixed use amongst the cases of the OfS standards of evidence self-assessment tool. In a couple of cases the tool had been applied in previous planning rounds but was not currently used. One person commented:

> *We haven't revisited that recently. But I definitely, definitely think they are useful. It's quite a big document, isn't it? Or a reasonable sort of wordy document. It's not the first thing I signpost people to, but I think it is really useful. Yeah, perhaps something that we could review again.*

3 THE TYPES/LEVELS DISCOURSE

The issue of whether or not it is appropriate to view the types of evidence as hierarchical has been a key aspect of the standards discourse. This Section discusses perspectives emerging from the research with case providers and the implications for the standards.

Key Findings

There has been an ongoing tension regarding whether a hierarchy of methods is implied in the standards: notably whether Randomised Control Trials (RCTs) should be considered a 'gold standard' for WP evaluations. Experimental methods are clearly a strong design when it comes to proving causality but all types of evidence provide very valuable knowledge to inform decision making on APP interventions in practice and there are a range of ways to increase the usefulness and validity of the results of all types of evidence to support decisions, as part of a proportionate approach.

Using the evidence standards encourages providers to move to increasingly robust approaches to using evidence, but the strength of evidence required for different types of decisions is a consideration. Impact evaluation complements other types of evaluation in evidence-based decision making and is used in a formative and summative sense as part of ongoing cycles of reflection and review.

Evidence that is layered over time to build new knowledge and move to successive levels of surety within each cycle is desirable and an opportunity to progress to stronger forms of evidence as the work continues and the amount of investment of APP resource grows. There is a tension in the current standards in the best way to move to stronger causal surety. The most expensive interventions warrant the strongest type of evidence, but these tend to be more complex and beyond the scope of RCTs which are perhaps best applied to discrete activities with direct outcomes.

Providers are looking to quasi-experimental approaches, especially where these are supported by developments in data possibilities (e.g. tracking systems and student records systems). Plus there is an interest in using theory-based evaluation and case-orientated methodologies.

It may not benefit continued efforts to improve the data and evidence base for identifying an impact if the strongest types of approaches are conceptualised in a purely experimental sense, without consideration of alternative interpretations, because experimentation in the strictest sense is seen as risky and only relevant for access and participation work in some exceptional cases.

3.1 Perspectives on the evidence types

The research suggests that whilst understanding of the strength of impact evidence and the claims that can be made as part of WP evaluations has likely increased as a result of the standards, there has been an ongoing tension in terms of whether RCTs should be considered a 'gold standard' for WP evaluations. It is perhaps fair to say that most of those involved in the research had engaged with the standards without necessarily agreeing that a Type 3 experimental design should be an end goal for evaluation strengthening when it comes to deciding what evaluation would be applied on the ground in different intervention contexts.

The potential for confusion on whether the types of evidence imply a hierarchy does not appear to have been helped by some perceived mixed messaging. Interviews with key informants suggested that there was a perception of policy emphasis favouring experimental designs (identified for example in the way evidence is assessed as part of meta-reviews and in terms of funding allocations to different types of collaborative studies).

> *I think that kind of idea of what works network and how other evaluations have been implemented across that network kind of takes us down a bit of a path and it can feel sometimes quite a bit like tunnel vision.*

> *I certainly think that people felt like we should all be doing Type 3, the 'gold standard' RCT. I'm very careful with that institutionally, when I talk about different types of evaluation, what sort is appropriate, what's feasible and all that sort of stuff. I think if you boil it down to Type 1, Type 2, Type 3, that can build on that kind of institutional idea that one certain type is what we should all be aiming for and it's that gold standard or the best.*

A key problem with taking a hierarchical approach was that it reduces the scope for creativity and flexibility on the type of evaluation that could be achieved.

> *The standards give us a lot internally to say and commonality of language. If they over do that emphasis on Type 3 it becomes restrictive and problematic.*

> *I guess maybe one of the reasons why we perhaps don't refer back very much as a document now is because it's kind of a common language, it's kind of like a living part of evaluation within access and participation and there's a general shared understanding as well as the different types and what sort of claims you can make. But we do have to very carefully balance that with the idea of hierarchy and value based judgments on what is better or what is the right type of evaluation.*

Different types of evidence are useful to support different types of decisions, and there is a link to the stage of development of an intervention, summarised in Figure 3.1. The evidence needs are likely to change over the lifetime of a project or programme as more/less formative and summative evaluations becomes useful/possible. This is especially in terms of intensive, sustained and progressive interventions (Type 1 evaluation might continue to be fine for 'light touch' interventions which expend minimal resource and/or where the relationship with outcomes is indirect).

The standards of evidence can help with decision-making by relating different types of evidence to different types of decisions: recognising that more important decisions (e.g. to commit a lot of resources to or completely change what is being done) will need more reliable evidence of impact. Type 1 evidence might be used to stop an intervention since they'd be no point in continuing if the initial data on outcomes suggested only poor outcomes were being achieved. If the intervention was proving itself by getting good results over time it might be that other types of causal evidence would be needed to make a decision whether to continue one intervention over another. There might need to be more surety that it was the intervention that caused the impact if the institution was recommending to roll-out a pilot, or to recommend the practice to other providers. However, types of evidence below complete causality might be considered sufficient if the method was considered rigorous enough: e.g. the counterfactual evidence from the evaluation was supported by other forms of reliable evidence of success, and/or there was convincing evidence that the programme theory worked.

Figure 3.1: Evidence to support different kinds of decisions

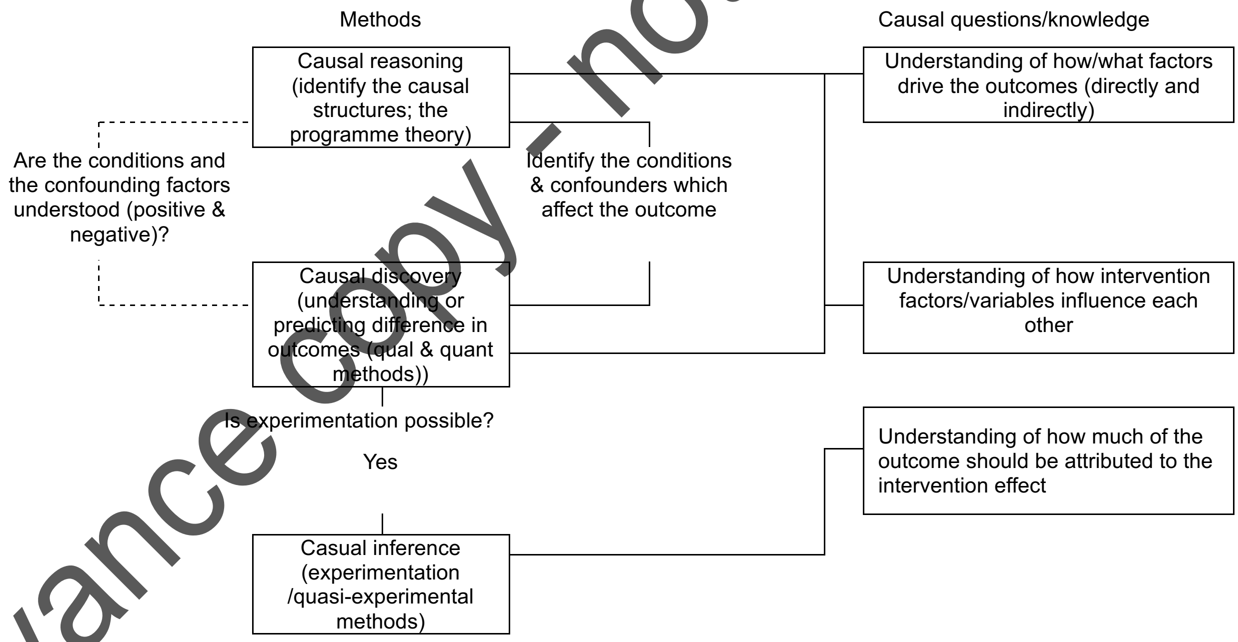
Decision:	Adopt it	Continue/Adapt it	Roll-out/Transfer it	Stop it
Useful impact evidence used	Existing external evidence to show intervention is associated with results, positive internal evidence of outcomes and practitioner expertise/ experience	Evidence to suggest there are positive results from the intervention compared to without it, even though causality isn't proven	Evidence to suggest that the investment in the intervention caused the desired result	No evidence of outcomes or evidence of negative effectiveness

The OfS requires providers to commit to at least Type 2 evidence for Intervention Strategies and the most resource intensive activities within the plans plus for new/pilot activities for which the contribution to knowledge will be greatest. However, there is a need for a considered approach so that the evaluation focuses on the key outcomes of most interest at the right time (which depends on phasing) and in turn will be influenced by the nature of the outcomes and impacts being evaluated. Where the shift to a type of evidence happens could vary. For example, an impact evaluation which embeds a pre-post evaluation design might embed Type 2 from the start (to provide a before/after counterfactual). A Type 2 evaluation which uses non-participant data as a comparison might need to look at the evidence at the end of a delivery cycle, and perhaps the Type 2 evidence could be strengthened over time by working towards a difference-in-difference approach, or by framing the evaluation as part of an experimental (Type 3) design. The above types of design could fit within a theory-based evaluation - in which case the current standards of evidence typology might hold. However, it is less clear where alternative (non-experimental) approaches to evaluation strengthening through theory-based methods fit into the standards framework (e.g. for example case-based and contextualised approaches or a contribution analysis).

> *I think the key challenge really is that the standards of evidence don't actually spell out evidence quality criteria. I thus worry that by conflating methods with quality criteria, we both uncritically accept 'gold standard' methods, and at the same time dismiss relevant causal evidence from non-gold standard methods.*
>
> *One of my key concerns is if it is Type 3 everybody just nods along, not questioning whether the method was applied appropriately, whether the information it provides actually answers the question we want to answer, whether samples were sufficient etc.*
>
> *From my perspective* [HE outreach] *it's very hard to put causality on it – almost impossible. Pressure on that is not helpful to us. Plus within the team it could undermine some of the evaluation that we've got established and been doing for a long time and the huge lot of evidence we're collecting.*

There is an interest in using theory-based evaluation and case-orientated methodologies as part of mixed methods evaluations. This type of approach to impact evaluation integrates causal reasoning and causal inference methods designed to test the programme theory underpinning interventions (Figure 3.2).

Figure 3.2: Causal Questions and Evaluation Methods


Methods
Causal questions/knowledge
Causal reasoning (identify the causal structures; the programme theory)
Understanding of how/what factors drive the outcomes (directly and indirectly)
Are the conditions and the confounding factors understood (positive & negative)?
Identify the conditions & confounders which affect the outcome
Causal discovery (understanding or predicting difference in outcomes (qual & quant methods))
Understanding of how intervention factors/variables influence each other
Is experimentation possible?
Yes
Understanding of how much of the outcome should be attributed to the intervention effect
Casual inference (experimentation /quasi-experimental methods)

On the ground the decision on the 'best' type of evidence is usually contextualised within the specific objectives and delivery context in place and the evidence needs. What counts as the 'best' design might be highly dependent on the nature of the question in hand. Often evidence-based decisions draw on a common sense interpretation of the strength of evidence in order to make conclusions and recommendations – especially for short term outcomes (for example on whether a skills development activity is supporting students to developed the skills required). Plus, isolating a causal impact is not necessarily the only purpose of evaluation and knowledge development in the higher education context.

3.2 Challenges to using experimental methods

There are multiple reasons why a purely experimental design can prove challenging in higher education contexts (Table 3.1). Apart from anything else the complexity of many widening participation interventions, particularly multifaceted interventions and strategies which cut across institutional functions, makes experimentation questionable because the designs are limited in what they can say about complex social fields.

> *I understand the concept of what good and relevant research is but from my perspective I find it difficult to do gold standard research*
>
> *RCTs aren't the reality in the situation we're dealing with.*

There is a perceived risk in committing scarce evaluation resources and capacity to an RCT because of the potential for unhelpful inconclusive results. RCTs provide the most compelling argument for causation, but they do this by disproving the null hypothesis. Inconclusive results therefore are hard to interpret. Depending on the numbers involved and the nature of the intervention (and in consideration of the relationship between the intervention and the outcome/impact and the existence of extraneous factors) there were concerns that RCTs were a waste of (scarce) evaluation resources.

> *How do you interpret* [inconclusive] *results? To be useful, you need a result because otherwise we'll just be like, oh, we've got no evidence that it's effective. The inference people make is that it's not effective because there's a null result. Not everybody who is engaging with evaluation perhaps has that understanding of how to interpret these kinds of results and what it means for them within their institutional context. So I think we've got to be really mindful of that in how we can both develop guidance but also present our evidence.*

Table 3.1: Potential strengths and weakness of randomised control trials in higher education settings

Strengths	Weaknesses
Seeks to define objective theories Tests a hypothesis Randomisation controls for extraneous variables Randomisation reduces biases by removing selection effects and controlling for confounding variables[10]	Time and resource intensive Relies on statistics (useless where statistics are not available e.g. small cohorts) Cannot deal with singular events (e.g. the specific personal effect an individual might say the intervention had for them) Can hit problems in identifying a control group (e.g. where there are ethical objections to randomisation when it comes to students facing disadvantages; can be issues with access to control group data) Probability that the results may not be conclusive Potentially low external validity (due to tight control conditions) May be of limited value in the assessment of long term outcomes because RCTs are frequently small and/or of too short duration for longer-term outcomes to be detected Might be danger of focusing on 'measurables' above other meaningful outcomes.

It is perhaps fair to say that in general the standards have pushed providers to make sure they have evidence to reasonably conclude that their interventions are likely to be making a material difference compared to what would have happened without the intervention (or not). From a provider perspective, having assurance that

[10] Randomisation doesn't completely remove biases because of the potential for differential dropout, measurement biases etc. Plus, in small groups, randomisation might not actually achieve balance because of the potential for uneven distribution.

the evaluation effort will evidence the worth of an interventions was probably more important than making claims to causality for exceptional interventions with the conditions for tight experimental control.

Less than half (three) cases in the sample had stated APP aims for a Type 3 evaluation. These cases were aiming to use a quasi-experimental design and included evaluation to assess: degree outcomes for students who utilised a contextual grade reduction at admissions; employment outcomes for students who access placement opportunities; the confidence, self-efficacy and continuation benefits of transition support; and continuation and future success outcomes of students participating in mentoring (amongst other things). Another case had included the use of inferential statistics in an evaluation involving assessment of cognitive strategies, metacognitive strategies and sense of institutional support/gains for students taking up different support and workshops compared to non-participants (using matched comparison). Key informants from two other cases noted they may consider if Type 3 evidence could be generated in future (subject to data availability).

There is a strong argument to be made that quasi-experimental designs are a better design than RCTs for evaluating student outcomes in an educational context. Firstly, the main use for RCTs is to provide a control based on randomisation so that evaluators don't have to worry about biases and extraneous factors. However, if evaluators knew enough about the problem being addressed to control well and had enough high quality data which allowed them to craft a good comparison group, then using a quasi-experimental method to craft a very good control might be possible to elicit a very strong impact evaluation design. Secondly, quasi-experimental designs might also be more fit for purpose in terms of what the evaluation is trying to find out. For example, there is scope for looking at how engagement in multiple activities affects the results, and/or issues of dosage.

> *As well as being resource/time intensive, you need a research question* [for RCTs] *so specific it's a challenge to get useful evaluation. The results of QED type evaluation are more retrospective, we need more flex there than just to test one hypothesis.*

The potential for a difference in how different types of quasi-experimental evaluations and other quantitative inferential designs (e.g. regression analysis) might be categorised against the standards was highlighted in the research (as a specific issue to be resolved). The size of the dataset is a key consideration in the methods that should be applied, raising additional questions about the most appropriate designs for small cohort interventions. This conclusion reflects other research findings into the efficacy of different kinds of methods, including a TASO-funded study which concluded that evaluation approaches specifically designed for small samples would be of higher relevance than experimental approaches.[11] Suggestions for alternatives included: theory-based approaches, contribution analysis, process tracing, or large-scale, deeper, and potentially mixed-methods process evaluation that explores every mechanism in the intervention's theory of change with a diverse range of students, including those who choose not to engage with the offer (p.2). The study points out the differences between focusing on the distant outcome (in the specific case these were progression to graduate jobs) and capturing the intermediate outcomes (in the specific case these were around self-belief, self-advocacy and employment-seeking practices). Taking different kinds of evidence together might allow for deeper understanding.

Furthermore, comments from the field highlight the need to combine different approaches to evaluation that maximise the potential for learning about why things work as well as what works. Understanding the why as well as the what was implicit in comments from key informants of what counts as good impact evaluation.

> *Even if we do this* [RCT] *actually the implementation and process side is so important because if we get a null result, we want to at least understand how and why. A stand-alone causal evaluation could actually be really unhelpful because it makes it look like something doesn't work. But why doesn't it work? Is it to do with the context? Is it to do with how it's implemented? Is it to do with challenges along the way? the timing? All that stuff is so important and that knowledge is what we need to generate.*

Overall, the 'small steps' approach to evaluation (Harrison and Waller, 2017)[12] is perhaps a more intuitive approach to evaluation evidence strengthening amongst practitioners in the field. Key informants said they

[11] TASO (2023) Efficacy Pilot Evaluation Report: London School of Economics' Disabled Students Career Appointments, https://taso.org.uk/wp-content/uploads/2023-10_TASO_LSE_Disabled-students-career-appointments-Efficacy_Pilot_Report_2023.pdf

[12] Harrison, N and Waller, R (2017) 'Evaluating outreach activities: overcoming challenges through a realist "small steps" approach', Perspectives: Policy and Practice in Higher Education, 21:2-3, pp.81-87. DOI: 10.1080/13603108.2016.1256353.

would like to see more practical guidance on how to improve within the type of evaluation that they are already doing. More practical help and support for evaluation strengthening within each type of evaluation was a strong recommendation from key informants.

> *It's the idea of you're better doing something and doing it very well than doing something that's too ambitious and failing. Yeah, because I'm really worried if we ineffectively run RCTs it's going to do nothing for anything. It's not really useful, whereas, say, a survey where people are really focused on ensuring the response rate and meaningful analysis and they've sampled appropriately and all of that kind of stuff, so little pockets of knowledge and building that capacity, I think are really important and I think that's where there's a gap in the sector.*

Clarity of what counts as good evidence to back up claims is especially important given the focus on external publication of evidence. Understanding that weaker and stronger designs are possible within each type of evidence is important as a tool for evaluation strengthening. As conceived within the original standards guidance, the framework does not see the evidence types as separate and discrete. The typology is designed to be cumulative and complementary as part of a layered approach to building the evidence base in order to support evidence-based decision making that would proceed through cycles over time and would complement other types of evidence from process evaluation, practitioner expertise and insights, and insights from research and theory.

Having a clear common sector-wide consensus by which to understand the 'strongest' impact evaluations is desirable; how this is conceptualised is central to the ways in which APP evaluations might inform sector-wide knowledge development as well as for localised decision-making going forward (discussed in Section 5).

3.3 Complementarity of impact and process evidence

How process and impact evaluation feed into each other as part of evidence-informed decision making in practice was a theme emerging from some key informants; with the suggestion that the knowledge and understanding built up on the implementation and process side was not only important to contextualise the results, but when used together with impact evidence added to the strength of evidence on which to base decisions. The question then for guidance on using the standards of evidence is whether evidence from implementation and process evaluation is supported within the definition of Type 1 evidence. This type of underpinning knowledge can play a key role in contributing to the narrative that is being built up about what works. This type of evidence also informs the interpretation of the results of subsequent types of evidence (for example being able to test the fidelity of the implementation and delivery approach when assessing the impacts of a programme).

> *You can't ever get away from process [evaluation] because it has an impact on the impact. How they [students] experience the activity.*
>
> *Yes it's a challenge to make claims on a small number of respondents and a focus group, but it's still useful learning. Of course you need to be careful on conclusions, and it's contextualised. Standards are valorising a difference – but what about use of evaluation in context – making our teams more efficient and impactful – redirecting resource to where they're more efficient.*
>
> *Sufficient data to measure outcomes should be prioritised but if you're unlikely to yield that type of data then you need to look at what's being delivered and process monitoring. You'll not get data on self-efficacy [the outcome] if no one turned up – maybe for marketing reasons etc, or whatever. You still need to know how many other initiatives are being delivered, what's the saturation of opportunities, and to test that - using data – but that's not the type of outcomes that the standards imply.*

Certainly, the sense emerged from key informants that using different types of evidence in combination can boost the opportunity for knowledge creation and practice learning. More guidance would be welcomed on how impact evaluation can be supported by other evidence.

> *Actually understanding why something works is really important in a Type 3 evaluation. I think sometimes it does get lost and I don't think although they do run the IPE [implementation and process evaluation] alongside an RCT, some of their narrative leans towards interpreting IPE or as a Type 1 or as a narrative and yeah, that's something that could maybe do with some unpacking. So maybe, some kind of unpicking of how they can have overlap and how they interact with each other and the complementarity.*

4 THE EVOLVING CONTEXT FOR IMPACT EVALUATION

This section discusses trends in higher education and the context for access and participation and the implications for impact evaluation, which we define as the systematic assessment of whether an intervention has made a measurable difference in outcomes for students. Observations are made on the implications for evidence generation and potential areas needing further guidance.

The latest round of Access and Participation Plans (APPs) has agreed the strategies and activities providers will take to address risks to equality of opportunity. The risk based approach adopted in the APPs has put the focus on the institutional risks and the experience of students. APPs set out how the impacts - in terms of the benefits for individuals and groups in achieving equity outcomes - will be evaluated over a four year period, working towards institutional level accountability targets which are embedded in gaps analysis alongside intervention-specific outcomes and impacts. Within the Intervention Strategy approach there is more emphasis than perhaps previously on lining up the contribution of different strands of work to the overall longer term objective/target. Providers were asked to share their evaluation plans and indicate the type of evaluation they were planning to use based on the standards of evidence. Dissemination commitments are included in APPs, although there is no set expectation on when and how providers should publish their evaluation findings.

The context of the new APPs means that APP evaluation faces some new challenges - and opportunities - that were less at the forefront in 2017 when the standards of evidence guidance was produced. Dilemmas for impact evaluation are identified below. The findings are based on the research with carefully selected key informants enabling general conclusions to be drawn about relevant trends within the sector.

Key findings

Developments in the context and wider and trends in higher education for access and participation work which came into place with the new APPs, have implications for the current standards of evidence in terms of: i) constraints and opportunities for certain types of evaluation; and ii) the kind of impact evidence that providers are seeking to identify impact and to underpin evidence-informed decisions.

The standards of evidence are concerned with whether outcomes can be attributed to an intervention (i.e. ascription of a causal link), rather than for assessing a contribution within the bigger picture. The issue of contribution not attribution is coming more to the fore Institutional Strategies are bringing together a suite of interventions to address different aspects of institution-wide problems. While this co-ordination is to be welcomed it is more challenging to isolate the impact of access and participation activities as they are becoming embedded in the day to day business of higher education.

The need to look at structural changes as part of sustained and embedded responses to addressing institutional risks, coupled with concern to understand and improve the student experience, may require new methodologies for impact evaluation. Alternative approaches are needed to address the challenge of demonstrating causal impact when outcomes are several degrees of separation from the delivered activities, and where outcomes arise from complex interventions, especially when they include a mix of direct interventions, embedded support, inclusive strategies and policy making.

The challenges include: dealing with evaluation of complex programmes; getting clarity on the purpose of impact evaluation knowledge development; linking short, medium and long term impacts; identifying methods appropriate to the context; understanding the best way to deal with the contextual factors in an evaluation; embedding evaluation across whole systems; using methodologies to involve students in evaluations.

A key finding was the trade-off between precise unbiased answers to narrow questions of impact and more uncertain answers to complex questions. The argument being that to be most useful to inform decision making in the current higher education climate evaluators might need broader perspective than answering causal questions with greater certainty.

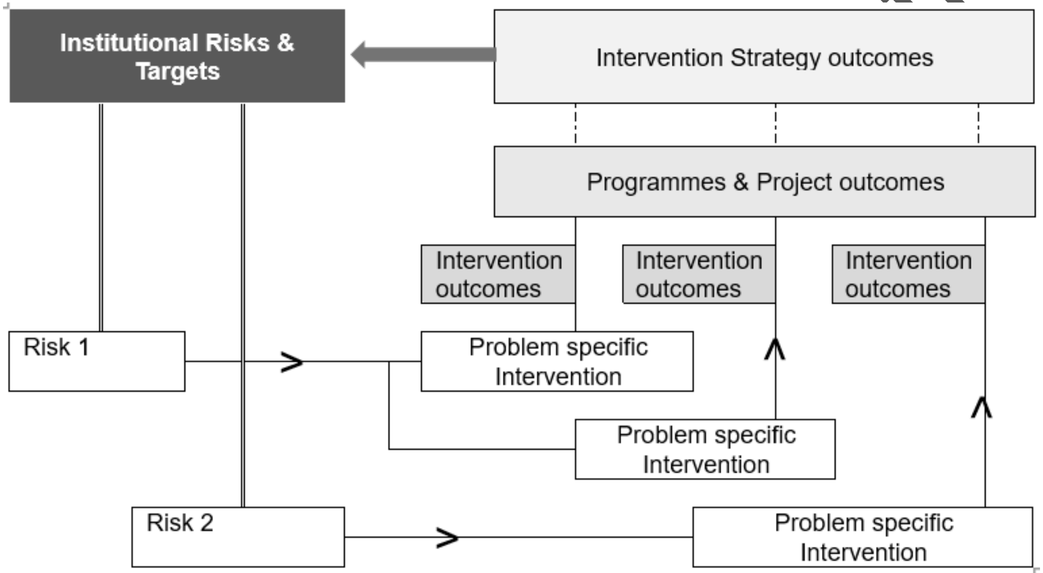
There is a danger that the standards of evidence will become unfit for purpose unless they can support providers in the collection and use of evidence in the evolving higher education context. Key questions to be addressed include: what counts as impact (contribution as well as attribution); the types of impact evidence that are most useful to support institutional decision-making; and how impact evidence is used alongside other types of evidence as part of cycles of reflection and review.

The standards of evidence are based around the kind of claims that providers can make when backing up assertions of impact. The methodologies captured in different types of claims should be appropriate to all contexts and access and participation work (in order to have general relevance rather than amongst those looking to do certain evaluation designs). More guidance is needed on how alternative methods of evidence generation required for complex types of access and participation interventions sit within the framework and what judgement can be made in terms of impact claims.

4.1 Risk-based interventions strategies

The risk-based approach within Intervention Strategies has brought together project specific interventions into a framework where there is more emphasis on looking at the outcomes and impacts with reference to the overall target. Impact evaluations at a project and programme level ideally need to inform understanding of what's being achieved overall. So an obvious question when it comes to making decisions on how to evaluate impact is what outcome or impact to focus on? Outcomes can be identified at different points (Figure 4.1). If the aim is to make claims of causality through an experimental design then there is a tendency to focus on programme specific intervention outcomes – which often tend to be stepping stones to other outcomes. Beyond this, the existence of parallel initiatives and the number of confounding factors may make absolute surety of causality impossible.

Figure 4.1: Kinds of outcomes (also note a link to short, medium and long term outcomes)

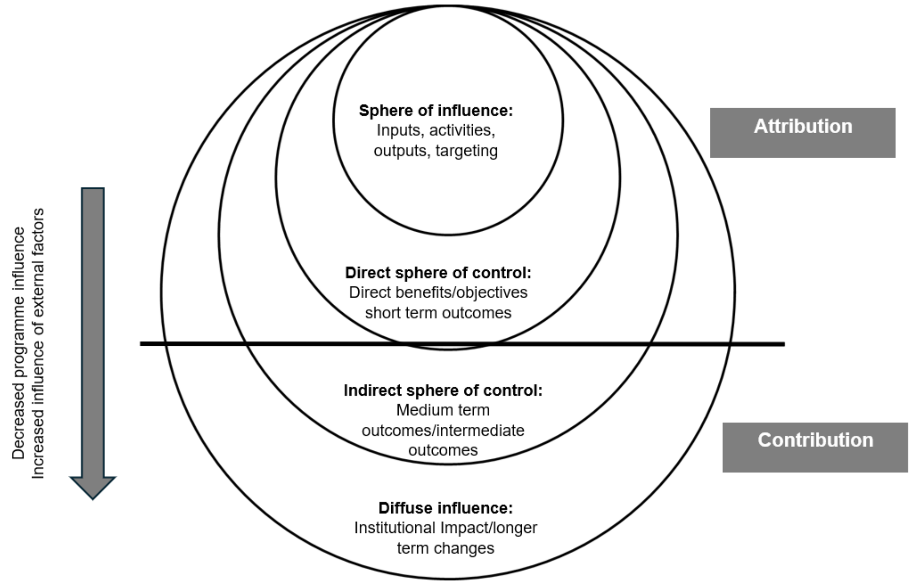

Institutional Risks & Targets
Intervention Strategy outcomes
Programmes & Project outcomes
Intervention outcomes
Intervention outcomes
Intervention outcomes
Risk 1
Problem specific Intervention
Problem specific Intervention
Risk 2
Problem specific Intervention

Access and participation work increasingly involves a high degree of complexity. Factors include: access to data and associated procedures, delivery across multiple teams (or sites). collaborative delivery, and community-based provision. Other factors affecting the outcomes could be inside and outside of higher education (for example, additional academic support outside the organisation because of a supportive family environment). In claiming a positive impact, higher education providers can struggle to determine how much of a particular intervention contributed to outcomes. Some interventions are designed to work in combinations with others, and some interventions are likely to have an indirect rather than a direct impact.

> *It can be hard to grasp what needs evaluating. There's the targets - we have a lot of general monitoring. It's easier now because the flagship intervention strategies provide structure. Programme evaluation is informing institutional evaluation, not just monitoring: How effective is the institution themselves?*
>
> *The challenge is very disparate. Take awarding gaps – a lot goes into it. Building an inclusive environment – halls, governance, workshops, programmes – disparate activities. It's difficult to identify outcomes other than gaps.*

Usually, a theory of change will include a series of intermediate effects that are necessary for the programme to be impactful and can test whether the assumptions about the final outcome held or whether implementation of the programme went off course or the activities had unintended consequences that countered the estimated effect (or other unforeseen benefits). Testing hypotheses, either with existing data and evidence or going back to collect more data (quantitative or qualitative) from stakeholders might be necessary, as well as considering running additional evaluations to explicitly test the emerging hypotheses. The precise attribution of outcomes becomes especially challenging when the evaluation is trying to capture the difference made to results over time which are increasingly out of the direct control of the programme (see Figure 4.2).

Figure 4.2: Attribution and contribution


Sphere of influence:
Inputs, activities, outputs, targeting
Direct sphere of control:
Direct benefits/objectives short term outcomes
Indirect sphere of control:
Medium term outcomes/intermediate outcomes
Diffuse influence:
Institutional Impact/longer term changes
Decreased programme influence
Increased influence of external factors
Attribution
Contribution

Causal and correlational impact evaluations tend to work more reliably where there is a fairly direct link between cause and effect, and where the number of confounding extraneous factors that are potentially going to influence the outcome can be better known and controlled for. However, it would not be a good result if all the work to develop the strongest impact evidence was only focused on measuring intervention outcomes for which the programme has more control. Accounting for the changes that happen over time in APP work can mean therefore that the confounding factors need to be embraced, understood, and part of the evaluation approach to explaining how causality works.

Balancing understanding of what works in terms of the ultimate goal and what works in terms of short-term outcomes requires layering up the evidence over time. Having longer timescales for impact evaluation in the new APPs is positive – it opens up the way for more longitudinal studies. In relation to HE outreach for example, it can take 4-5 years to know if an intervention in Year 9-10 affected university take-up (taking account of not only student journeys through education but also the data lags in the system). Longitudinal evaluation requires the systems to be in place for this.

However, the key informants were cognisant of practical difficulties for evaluation over longer cycles where activities are changing and evolving on the ground, including the potential of moving away from the original blueprint or hypothesis being tested. In this context a key dilemma was whether to set up trials designed to test the programme theory – which could take years to get the results and inform decisions - or to embrace a developmental approach to learning from evidence which involves integrating emerging evidence into practice – which loses the fidelity of the model being tested.

> *There could be discussion between staff and students which changes what they're doing. Thing's that practitioners do independently can shift a project, If there's a change of approach, the evaluation automatically changes.*
>
> *A challenge for APP* [evaluation] *is that we're working with two mindsets. In the university's mind that years' gone, move on, not thinking about the implications* [for evaluating the programme theory]. *The university can forget we did a project – doesn't always scrutinise – despite the fact they're some cool exciting things going on.*

Challenge 1: How to deal with complex programmes

Evaluators working across Institutional Strategies and the student lifecycle are having to embrace complexity in the approaches they are developing. This could include thinking about issues of contribution as well as attribution. In practice providers may be assessing claims to impact on the basis of judgements of reasonableness and 'good enough' evidence in our context. This is fine up to a point but might be undesirable because the basis of decisions become less transparent, which goes against one of the key objectives of the standards of evidence. A key question is where does evidence from complex and contextualised evaluation contexts fit within the framework? How this type of evidence should be judged is difficult and depends on whether qualitative tests of reasonableness, credibility and coherence of argument, commitment and rigour and so on are as considered valid an comparable to quantitative tests of validity, reliability and objectivity. Whilst, the current formulation of the standards embraces the use different kinds of data, the methodologies are as yet fairly underdeveloped especially for theory-based evidence.

Challenge 2: Purpose of impact evidence in relation to different types of learning

The use of impact evidence in evaluation can conflict with a continuous improvement approach. On the one hand there is benefit from access and participation practitioners engaging in a process of reflexive practice and review – which is a tenet of educational practice – and undertaking adaptions as they go along. On the other hand fairly fixed models of delivery that can be systematically tested are desirable in order to assess the results of the model. Having systems in place that can capture differences made over time appear to be important in both cases, along with understanding of the causal mechanisms. However, the above approaches could take evaluators down different evidence pathways. The learning about a fixed model of practice is the recommendation to continue (or not) but there might be limited understanding in a fixed approach, if the initial evidence suggests it’s not effective. The learning from evidence of the implications of practice changes is the practice improvement itself (an action research approach).

Challenge 3: Linking up short, intermediate and long term outcomes/impacts

There are knowledge gaps in higher education about the linkages between short/medium term outcomes and final impact indicators which might require new types of evaluations, perhaps involving modelling. For example, is it enough to say the intervention raised study skills in higher education without understanding the relationship between study skills and actual HE achievement for different groups of students, and the implications for attainment gaps? This aspect could perhaps benefit from a sector level focus in order to assist evaluators in interpreting impact results from proxy indicators. For example, the extent to which results from the Assess and Success Questionnaire (ASQ) are predictors of student outcomes. Further insights into using short and long cycle evaluation approaches could also be helpful in this respect.

4.2 The delivery context

The potential for using different methods of evaluation depends on the context - the size of the provision is an obvious one, affecting access to datasets, and the types of data linked to the intervention. Small and specialist providers may be less able to engage in quantitative method of evaluation. They may be working with more resource intensive methods and may need more practical support and guidance on steps that can be taken to evaluate in their context.

> *It's capacity and also the nature of our institution. A lot of our interventions and evaluations are sort of small.*
>
> *Because we're a small specialist institution, we don't have a lot of particularly targeted interventions from the APP. We're not having thousands of students participating. We can, create some comparative group but there are few opportunities for us just because of the nature. Yeah, who we are, the types of interventions we deliver. And yeah, evaluation capacity as well.*

Both quantitative and qualitative approaches to impact evaluations are encompassed within the standards of evidence but qualitative methodologies are underdeveloped in an applied sense (e.g. there are few examples of use of qualitative evaluation techniques such as case study methodologies or Qualitative Comparative Analysis (QCA)). There is a tension because qualitative methods can be more time and resource intensive (when compared with use of surveys for example) making them a bigger resource requirement, even though the scale of provision is smaller.

Small cohorts exist in all sizes of providers, so the problem is delineated by the delivery context as much as the size of provider per se.

> *Thinking about samples, our outreach work is very small, very targeted. We might be working with 12 students over a period of two years.*

4.3 Nature of the outcomes and impacts

There is a general trend to embed evaluation into intervention delivery mechanisms in order to make sure data for evaluation is being generated appropriately from the ground up. This potentially brings other factors into play, such as wider access and more data, the need to build relationships between teams, and to build evaluation into practitioners day to day roles. These factors affect what approach to evaluation is practicable, and at what level (for example outcomes measured across programmes rather than for each intervention within a programme).

> *Some projects are too small to* [evaluate] *ad hoc – it's a matter of establishing and embedding evaluation into service outcomes monitoring.*

> *We made huge improvement in embedding evaluation into activity. Projects funded in APP, core services, are embedding evaluation. Our dashboard portfolio programme keeps a view of all the work that takes place.*

> *We're evaluating a racially inclusive curriculum assessment piece of work, that's been delivered to all students on our undergraduate programme in the first year: that's 1000 students all leaving this intervention. So actually that is where we have to go to maybe look at Type 2 and Type 3.*

The potential trade-off in getting more data for evaluation is making sure the impact measures are relevant to the delivery context (since there may be a danger of the measures being specified at too high a level to be meaningful to the activities being delivered).

> *It can be hard because there's no connectivity, we can't compare activities, but a lot should be addressing a particular problem. Comparisons are difficult because the benefits might not be similar – but it's essentially one project (objective) looked at from different angles.*

Depending on the specifics of the intervention, the desired outcomes and impacts will affect measurement in a range of ways, as well as the research question. Impact evaluation which aims to explore how sense of belonging has changed over time requires a different approach to one which seeks to determine the difference made to HE continuation rates (and both might be needed as part of a rounded approach to impact evaluation). APPs usually cover a varied set of objectives, which could be captured in various ways, and some things are harder to quantitatively evaluate than others.

> *Our intervention strategies are on risk and some of those are institutional. So for example, an institutional risk of replication of inequalities of the professions. So we're looking at how the profession is very white, very middle class and we risk replicating that because we have differential outcomes. That covers absolutely everything from outreach work in schools and representation through to activities around recruitment of staff. The long term impact is hopefully to see a more diverse pool of people entering the profession and progressing through to senior level. We can contribute to that, but our strategy isn't solely going to resolve the structural issues so I would find that very difficult to evaluate.*

The ability to control for the external factors will also vary depending on the nature of the outcomes being evaluated (and the timescales involved).

> *When we think about outreach, we have less knowledge about potential other factors that could impact the outcomes for those students. Whereas when we think about success, we have, some knowledge about what other factors are in play institutionally. That helps, but it's really difficult to control for in an experimental way because students might self-select to access student support alongside our other interventions.*

The delivery factors increasingly need to be viewed as part of the evaluation since a delivery context may not have perfect control over the intervention - understanding the context for delivery helps to explain what happened. The traditional approach to randomisation or a quasi-experiment is designed to eliminate differences, but theory-based methods of evaluation can usually better deal with evaluating complexity. Unpicking the theory can help to pin down what has happened in terms of the causal linkages and underpinning mechanisms.

Challenge 4: Methodologies appropriate to the context

There is increasing interest in theory-based evaluation as a way of maximising the learning from evaluation (and making best use of evaluation resources). This is not only for ‘small n’ providers and delivery cohorts (who also have less scope for quantitative methods so have to be more creative in their approaches to evaluation) but also for other types of provision given the complexity of access and participation interventions in general. An issue for the standards is providing a framework that is even handed in terms of ability of different types of providers, and different delivery contexts, to engage with stronger types of evidence within Type 2/3 evaluations.

Challenge 5: Dealing with context in impact evaluation

A central question in terms of which methods are most useful for impact evaluation is how to deal with contextual factors as part of an impact evaluation: should context be controlled for (e.g. as in an experimental approach) or should the contextual factors be embraced as part of the explanatory factors for impact (e.g. as in a realist approach, or theory of change approach, to evaluating impact). Evaluation of complex interventions is likely to require a contextualised understanding of the mechanisms involved, and this can also be the basis of making causal inferences.

4.4 Organisational/structural changes and the ‘whole provider approach’ (WPA)

OfS guidance defines the whole provider approach as ‘a description of how staff from departments and services across the provider are led and engaged to ensure that its students are supported to access, succeed in and progress from their time at the provider’.

The WPA raises questions regarding aspects such as leadership, culture and inclusiveness.

> *For me WPA is everyone understanding the strategy and that we’ve got gaps: generally people having an understanding of the barriers. A change of culture of the place.*

Organisational and structural changes tend to be qualitative and values-based which is a problem for the application of quantitative methods. Moreover, developments in a whole provider approach including joining-up of different services to meet an equity objective (with linkages to quality mechanisms (i.e. The Teaching Excellence Framework (TEF) and B3 accountability metrics)). The fact that widening participation has moved out of the realm of ‘special’ interventions is another reason why new lenses through which to look at impact are potentially required, because the delivery context becomes even more complicated when thinking about institutional level impact. One of the main implications emerging for impact is disentangling the contribution of different interventions, programmes and initiatives and the regular business of inclusive higher education.

> *As a strategy, I can see what the vision is and how it will contribute, but we can't just run some kind of quasi experimental evaluation and say, oh, is our intervention strategy on inequalities effective?*

> *There's lots of activity going on either at broad institutional level or departmental level. We're all trying to address the same problems. How can we over here say our racially inclusive assessment piece is the thing that's improved outcomes for attainment for Black and Asian students when student support might be saying, well, actually we've made our coaching model more racially inclusive, for example. That’s why it’s important to know how and why things are working and not just throw a load of stuff at a wall and see what sticks.*

This context is different to the one in which the standards of evidence were originally developed (which was the narrow context of thinking about the impact of outreach on aspirations for higher education).

> *Obviously there are still people who are very outreach focused within their work, but most people I guess within the space would be looking at that cross lifecycle aspect. If you look at it from the point of view of an intervention strategy, you might have organisational indicators and you might have staff related outcomes as well, so it's not just student outcomes that you might be interested in. It's hard to do an RCT on a staff outcome.*

Case institution approaches to whole organisation level evaluation are not yet fully formed. Overall, APP evaluations remains intervention focused, and there is more work needed on evaluation types and methods that will be most appropriate for institution-wide evaluations.

> *We don’t have in-house evaluation capacity for understanding ourselves. For understanding how all our parts work towards the end goal.*

> [Evaluation is] *easier if there's a team in a space delivering on a plan. Shifting to a whole provider approach has added a layer of complexity. Can existing mechanisms here change and be made fit for purpose? Stuffs happening across the board but not all of it is being evaluated. That relies on them* [practitioners] *being forthcoming. For WP we are focused, but trying to embed evaluation is harder in programmes where we don't know the impact and there's no* [evaluation] *resource. We see duplication, similar initiatives especially in race equality. We've all got a vested interest, but what it is doing? where does evaluation fit in bringing things together? What are the initiatives and how to evaluate them?*

Different strategies and initiatives are linking up to support the collective effort.

> *Tracking and benchmarking on B3 metrics – that's what everyone's looking at. Evaluation is part of that – outcomes and measures versus awarding gaps comparisons. It's opening the door to anything that will lead to an increase in B3 metrics and this is probably doing more than one thing at the same time.*

A helpful consequence for APP evaluations is leverage in the institution for evaluation strengthening – one example of this was evaluators being able to use student outcomes datasets for evaluation as well as monitoring purposes. In some case providers, evaluators are moving out of WP and becoming connected to overall strategy, or other core institutional functions – i.e. strategy and planning, data and analysis teams. Linkages to wider strategies helps to push forward APP evaluation activities because the evidence being collected benefits understanding beyond APP work and opens up new evidence possibilities.

However, at least one key informant felt unsure about how the organisation would meet the challenge of evaluating at a cross-institutional level in the future.

> *I think there are people now within institutions who will be able to deal with that rather than where we were six years ago. And it's good, obviously, that it's moved on, but what is going to be useful in the future as well when all these issues come through? Not just for APP but from a TEF point of view.*

Challenge 6: Embedding evaluation across whole systems

There is a challenge of embedding evaluation across whole systems – and this is especially the case in a large complex organisation that is constantly changing. Whole system evaluation is likely to take some time to achieve, since it requires both top down and bottom up commitment to the evaluation strategy and effective communication either way. Being able to use evidence for decision making on interventions to inform developments at a provider level argues for coherent strategies which pull together monitoring systems, implementation and process evaluation, and outcome and impact evidence tools, along with tracking/data systems and methods which triangulate different types of evidence.

4.5 Concern for student involvement

A focus in English HE on the student experience and student involvement is perhaps changing the way evidence is viewed: participants in this research for example indicated that evaluation which surfaces the student voice is perhaps the type of evidence which decision makers pay most attention to. There is also interest in co-creation aspects and more holistic approaches to understanding the lived experience as part of the evaluation effort. This issue has implications for the standards framework that is applied across all APP interventions. For example, experimental approaches seek to control for the individual effects, whereas there is increasing concern to treat everyone as an individual.

> *Why did they participate – everyone is a complex individual with lots of intersectional characteristics, you can't ignore all of that. Most institutions can't do that type of scientific trial – it's their* [the participants'] *one shot. Give everybody everything that can help them. We work on layers of evidence and more holistic approaches.*

> *You need to know who got the intervention. We tailor to the programme needs, rather than what problem we think should exist. There's no point doing good work is it's not fit for purpose.*

Practitioners and those making decisions on interventions are perhaps less inclined to fix the outcomes in advance, and keener to capture the needs, benefits and experienced as perceived by those taking part in interventions, including through more participatory forms of impact evaluation.

> *From an ethical social justice side – they want us to work with students. So we need outcomes harvesting - not just to do and learn – and the opportunity to reflect. Experimentation is out. That's not just epistemic, it's about having a reflective space and timeframe.*
>
> *Interestingly enough a distinction that* [the standards] *neatly miss on is whether we're addressing the right problem. We support collaborations with students - invited spaces – because we want to be as critical as can be. What are their challenges? Double-loop learning challenges values.*
>
> *We are balancing the need for giving as much power to students and meeting organisational needs.*
>
> *People's voices are important, stories are important - fundamental to what we're trying to do. What's the impact and benefits of initiatives for our students?*

Measures of student experience and related concepts (such as student engagement, sense of belonging) have come in scope as intermediate outcome measures (also influenced by a focus on causal mechanisms as well as outcomes). Cross-institutional tools (such as student surveys) which collect data from a wide group of people thereby set up the opportunity for comparative and correlational evaluation work, and the possibility of quasi-experimental designs. However, not everyone is sold on this type of approach:

> *It's about students' experience in this setting and them getting the most out of what we provide. At the centre is the student who studies with us and how they experience everything. You have to care about the students not the data. They're not just data, they are people, they're not doing it* [higher education] *for the sake of it: we need to understand what is their capacity to achieve.*

Challenge 7: Student involvement in impact evaluation

Researchers and practitioners acknowledge the importance of student engagement in evaluation and are seeking new methodologies for this (although overall evaluation processes requiring student involvement in quantitative studies tend to remain predominantly designed and controlled by staff (e.g. surveys)). Participatory and co-creation approaches are important to bring in new perspectives, but this kind of research tends to be exploratory and it is not clear how such methods support impact evaluation. There are potentially many benefits of engaging students in evaluation - including perspectives on what impact is. Collaboration in evaluation opens up dialogue on the kinds of outcomes and impacts that are more relevant to student themselves. There is a sense that evaluation practices can be augmented or redesigned to enhance student engagement, but this an underdeveloped area which is likely to have implications for APP evaluation and evidence.

4.6 Discussion

Impact evaluation in higher education has a dynamic perspective because higher education providers are complex and adaptive, which puts the focus on contextualised understanding of how and why an intervention generates change and the enablers and constraints on its delivery in each setting. Therefore, intermediate outcome measures capture changes in the system as well as for individuals/student groups (e.g. changes in relationships internally, normalisation of practices etc). The approaches need to be flexible enough to allow providers to prioritise, use information for decision making and select an optimal evaluation approach to answer their particular impact questions.

The sector probably needs to give more consideration to evaluating complex interventions. A key finding was the trade-off between precise unbiased answers to narrow questions of impact and more uncertain answers to complex questions. To be most useful to inform decision making in the current higher education climate evaluators might need broader perspective than answering causal questions with greater certainty. The providers in the sample were taking a comprehensive approach to evaluation: where questions about how something is working and the impact that this is having are asked in tandem as the work goes along. Then the evidence informs whether the intervention proceeds to the next phase, returns to a previous phase (i.e. a re-development of the programme theory), repeats, or stops.

The framework, commissioned by the Medical Research Council (MRC) and the National Institute for Health Research (NIHR)[13], may be a useful reference point when thinking about impact evaluation of complex interventions. This framework is designed to support complex intervention evaluation to maximise the

[13] https://www.journalslibrary.nihr.ac.uk/hta/HTA25570

efficiency, use, and impact of research in the health sector. The latest MRC guidance is an update which seeks to help researchers work with other stakeholders to identify the key questions about complex interventions, and to design and conduct research with a diversity of perspectives as well as an appropriate choice of methods. The guidance takes a holistic perspective of where impact evaluation sits in terms of asking a broader range of questions to inform evidence based decision making (e.g. identifying outcomes and impacts, theorising how things work, looking at how an intervention interacts with the context, the contribution to systems change, and the value for money issues). The MRC framework recognises the need for evaluation at different levels (from individual to societal levels) and supports making connections between them. Learning from this approach may support how the higher education system thinks about evaluation of complex interventions – certainly conceptualising higher education providers as complex systems potentially helps understanding of the interaction between an intervention and the context in which it is implemented.

5 IMPACT EVALUATION FOR EXTERNAL KNOWLEDGE GENERATION

A key objective of standards of evidence is to facilitate robust impact evaluation plans, so that quality evidence will be generated. There is increasing emphasis on bringing together evaluation for accountability with evaluation for sector-wide knowledge generation. Ultimately, by sharing findings about what is expected to work in what circumstances with whom, what is proven to work and what does not work, the use of standards were designed to help to ensure that access and participation activities and funds are directed to the most effective activities.

The APPs commit providers to publishing their evaluations. The launch of the Higher Education Evidence Library (HEEL) will further support evaluation evidence generation and sharing by building up a central repository. Thinking forward, there are implications for how evaluation for external consumption is conceptualised to support submission to the HEEL, how assessments/judgements are made on published evaluation reports, and the types of evidence that are needed to support replication/transferability once impact and effective practices have been identified. This Section discusses perspectives in the case providers and the implications for evaluation practices.

Key findings

The expectation of evaluation for external publication set in the APPs is blurring the lines between evaluation and research. Within institutions, there are advantages in conceptualising some types of ‘flagship’ evaluations which are designed to inform the external (as well as the internal audience) as research since it could encourage increased scrutiny, and more resources to undertake the evaluation and make better considered evaluation design choices. However, the amount of time/effort needed for external dissemination activities is unclear and there is a concern to make sure it is proportionate within an overall strategy so that evaluation most effectively supports delivery and outcomes for disadvantaged students.

There are advantages to reports in the HEEL being categorised against the standards to direct readers to how the knowledge can be used, but any assessments need to value the contribution from different evaluation contexts equally. However, there is a difference between the use of standards as an evaluation planning and development tool for APP evaluation work, and as a tool for retrospectively assessing the claims being made from specific impact evaluation reports. Using standards as a planning and development tool before/during an evaluation guides how the evaluation is conducted in order to ensure learning and improvement. Using standards as a retrospective assessment tool judges the validity of the causal evidence that has been produced in order to decide if an intervention is to be recommended more widely or not. If the distinction between these uses isn’t clear, there’s a risk that evaluation is judged unfairly (for example, criticising an approach for not meeting criteria it was never guided by). Overall, there is perhaps an appetite for sharing wide-ranging evaluation studies cutting across all types of evidence and applying value judgements of evaluations on their own merit or with reference methodological quality criteria rather than evidence standards..

Decision-makers tend to lean towards evidence that reflects their particular circumstances and concerns. Useful evidence to support replication and transferability of interventions that have impact and associated effective practices needs to be contextualised to draw out the implications for other providers. If the knowledge is to help with replication, then understanding of the implementation and process factors, including the targeting and deliver factors, need to be reported alongside the evidence of impact.

5.1 Generation of external evidence

The emphasis on external knowledge generation has added a new dimension into APP evaluation work.

> *If you're talking about evaluation then you're not contributing to knowledge in the same way - the traditional research principle – evaluation is for whether or not something you did had value and the implications for your practice – others can learn but that's not the primary objective. John Blake talks like it can be research – if we talk about evaluation as research – it can be confusing.*

The focus on external publication raises issues about whether evaluation should be prioritised with an external or internal audience in mind. The ideal is both. The type of outputs which are needed and amount of time and effort that access and participation colleagues will be expending on materials for submission to the HEEL is as yet unclear. It's early days but implications were identified for how APP evaluations which have an external audience in mind are viewed. A common approach involved prioritising certain flagship evaluations for a sector audience as part of an external dissemination strategy. This approach could be beneficial – for example by adding an additional level of scrutiny and review.

> *Confirmation bias is built in – there's a strong presumption it works. We don't challenge ourselves, unless we turn it into a special project.*
>
> *When we lean into the research zone that helps us. When there's a requirement for research experience and things move more towards research they move away from a localised feel of evaluating practice.*
>
> *We might pick individual projects at Types 2-3 – make them evaluation priorities and commit to publishing. It might not be super scientific, but reasonable, and interesting to the institution and regulator, and these might usually be newer, but broadly feasible. A novelty but important and what we know about.*

For this kind of dissemination, there was perhaps a sense that the types of evaluations that would be disseminated for sector-wide knowledge development should be viewed as 'evaluative research'. This would define the studies fairly broadly as a method used to assess the design, implementation, and outcomes of programmes/interventions. Impact evaluation studies could come within the overall banner but it would also include other objectives such as the effectiveness, efficiency, and relevance of interventions in achieving their intended goals. All these types of research are useful to measure performance and whether a particular intervention is working, or identify areas for improvement, making them useful for informed decision-making. Practitioners in institutions are perhaps also interested in understanding how students experience particular interventions.

> *Evaluation is research – you get a report – that legitimises it to the wider HE crowd. We do look at it like research. We're doing the same activities: a literature review; gathering data/evidence; making recommendations. You can't be snooty about it, we're researching the effectiveness of our programmes.*

A parallel approach was to go down the line of increased sharing of evidence of the overall impacts and benefits of what perhaps might be seen as 'the package' of interventions within an overall Institutional Strategy or programme. For example these could be reported as part of annual reports/reviews which are disseminated externally.

5.2 Making post hoc judgements about impact evidence

Where resources are short the decision perhaps needs to be based on the imperative to use evaluation for the best effect to improve effectiveness and outcomes for students. Resource constraints for evaluation external reporting could potentially discriminate against some types of organisation for which there is less interest already in external publication (for example within providers where there is less focus on the generation of academic articles). Key informants spoke about the need for wide-ranging kinds of outputs of evaluation evidence, emerging across all kinds of higher education provider contexts.

> *I think particularly as someone who started their higher education career in further education and knowing the capacity that those institutions have, they need to know that their evaluation and their kind of learning, however, they can generate that and present that still has value within the sector.*

Categorisation of disseminated evaluation reports in the HEEL repository according to standards was potentially considered a good thing in order to help readers assess the nature of the evaluation and how the findings should be used. However, some sense of trepidation was evident in relation to how evidence will be value judged.

> *We think we've done a good job of balancing OfS requirements and a decent standard. But what does it look like across the sector? Plus there is an issue about how we share individual evaluations with colleagues and the fear of how that's presented. We've the right expertise but a challenging lack of resources. There's that nervousness around sector.*
>
> *They need to meet people where they are and provide meaningful practical support and expectations and don't make people think that they're kind of contribution to knowledge isn't useful or isn't valued.*

The application of the evidence standards to the assessment of reports submitted to the Uni Connect evidence base as part of the Uni Connect national evaluation work was a - potentially unhelpful - precedent. The national evaluators identified where evaluation designs were judged to be weaker/stronger in relation to the research designs. Whilst from a methodological standpoint some designs generate less reliable evidence on which to base decision-making, weak/strong are pejorative terms that can be damaging to those involved, plus the judgement does not take account of contextual factors/constraints affecting the design choices being made.

> *I wasn't always keen on that weak Type 2 and those judgements that CFE used to make on evaluation for the Uni Connect, because it was really disheartening to be told that your evaluation was weak. So we need to think about language. But it's the idea of you're better doing something and doing it very well than doing something that's too ambitious and failing. An inconclusive RCT will do nothing.*

There are differences between the use of standards as applied as an evaluation planning and development tool for APP evaluation work in general, and as a tool for retrospectively assessing the claims being made from specific impact evaluation reports (although the current standards of evidence tend to be applied for both purposes). For evaluation planning, multiple considerations are at play – not least the aim of the evaluation in question – and the need to take into account what data is achievable in the context. Plus there are question of proportionality. At sector level it would not be helpful to have an approach that potentially discriminates certain types of providers or dis-incentives to share evaluations. The sense appearing from the fieldwork was there is a demand for all types of evidence which could contain 'nuggets' of knowledge about effectiveness in generating impact (as well as the impact itself). Conversations in the field revealed a preference for evaluation that supports practice improvement (i.e. as part of a developmental approach). This contrasts with the hypothesis-test-decide model of impact evaluation where results are only useful in periodic decision-making to either scrap or continue rather in the ongoing endeavour to adapt/improve.

One can speculate that the higher the perceived risk involved in the process of submitting evidence for external scrutiny the less engaged the sector will be in coming forward with their evidence. In turn this may limit the usefulness of disseminated evaluations – i.e. if they become confined to certain narrow types of evidence.

5.3 Evidence to support replication of knowledge and transferability of effective practice

Within the sample of cases included in this research there were differences in which types of external studies were used to inform practice: there were preferences for certain types of information to be considered more valuable for decision making than others (e.g. peer review studies and meta reviews over practice based outputs). Plus, different types of providers may have varying capacity for evidence collation. Some providers seemed to be most alert to evidence which surfaces the views and lived experiences of students.

> *It's good that we are seeing examples of diverse methods –and examples we can use (for example appreciative inquiry).*

There is an argument to be made that any taxonomies of evidence should be focused on making the evidence base as useful and usable as possible as knowledge that providers can transfer to their own contexts. 'Hard' evidence of impact is obviously most valuable in making decisions about 'what works' but there was also demand for contextualised learning and understanding 'how' and 'why' interventions might be effective. Evidence to inform decision-making in the case institutions could come from a range of sources but providers are seeking to understand the context in which the work took place, in order to assess the applicability to their own particular context and institutional needs.

> *The main criteria is does it fit/is it narrowly similar in terms of the intervention. Meta-level conclusions are OK but there needs to be an internal consistency argument based on the findings. Usually there's no opportunity to compare evaluations.*
>
> *I would say people here are suspicious of external expertise. The APP research could show other people tried it, but there's a reluctance because there's some unique qualities here. We want to develop contextualised evidence now. Projects which seemed like a good idea were not thought through. It's not just finding out why change has happened – but why is it so impactful at this institution?*

There are information needs when it comes to supporting the use of standards of evidence to help with the transferability of proven interventions to others in the sector. Evidence of causal impact is part of this, but it needs to include a contextualised understanding of what was delivered, with whom, in what context, and how. These questions becomes particularly crucial when thinking about practical transferability.

It is interesting to note that other standards frameworks which aim to support replicability go further in layering evidence types to continue to build the case for the positive impact potential of the intervention in question. For example, Levels 4 and 5 of the NESTA Standards of Evidence for Impact Investing are concerned with building independent replication evaluations of the conclusions to confirm the findings, and the development of manuals, systems and procedures to ensure consistent replication and positive impact.[14] This approach is designed to balance the need for evidence with the need for the transfer of innovation. It highlights that decision-makers are concerned to understand the conditions for transferability and the delivery aspects involved in proven interventions, as well as the results. The Standards of Evidence for Impact Investing as also explicit about the role of external evaluation/verification of the results, because this can add a further dimension in terms of reliability of the results of evaluation.

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

6.1.1 Implications of the standards of evidence

Overall there was a sense that the standards of evidence have positive benefits. Internally, they support development of evaluation strategies; decision making and approaches to doing evaluation. Externally they are helpful in communication to the regulator and sector colleagues. The standards have played a role in evaluation capability building and have built understanding of what counts as useful evidence for decision-making, pushing providers to implement increasingly robust approaches to assessing the effect of interventions and greater surety that interventions are making a difference. The standards support the decision making processes at project, programme and institutional levels. They support planning and have contributed to good practice approaches particularly in relation to building the rationale for interventions and obtaining clarity of outcomes and impacts using theory of change.

The standards provide a 'common language' and a 'touchstone' for the sector on evaluation. However, there is also a risk that conceptualisation of the standards tends to be unsophisticated, which may limit the extent to which they can be practically useful for the purpose of further impact evidence strengthening across the sector. There is a danger of the standards becoming diluted and a 'shorthand' for a methodological distinction between more quantitative or more qualitative approaches to impact evaluation and evaluation in general.

6.1.2 The Types/Levels Debate

There has been an ongoing tension regarding whether a hierarchy of methods is implied in the standards: notably whether RCTs should be considered a 'gold standard' for widening participation evaluations. Experimental methods are clearly a strong design when it comes to proving causality but are less useful in explaining the reasons for a particular outcome. Complementary types of evidence provide very valuable knowledge to inform decision making on APP interventions in practice. In addition, there is a range of ways to increase the usefulness and validity all types of evidence to support decisions, as part of a proportionate approach.

The evidence standards encourages providers to use increasingly robust methods of generating evidence of impact, but the strength of causal evidence required for different types of decisions is a consideration for what

[14] Puttick, R. and Ludlow, J. (2012) 'Standards of Evidence for Impact Investing.' London: Nesta.

type of evidence is most useful. Impact evaluation complements other types of evaluation in evidence-based decision making and is being used in a formative and summative sense as part of ongoing cycles of reflection and review.

Evidence that is layered over time to build the knowledge and move to successive levels of surety within each cycle is desirable and an opportunity to progress to stronger forms of evidence as the work continues and the amount of investment of APP resource grows. There is a tension in the current standards in the best way to move to stronger evidence of causality. The most expensive interventions warrant the strongest type of evidence, but these tend to be more complex and beyond the scope of RCTs which are perhaps best applied to discrete activities and their direct outcomes.

Providers are looking to quasi-experimental approaches, especially where these are supported by developments in data possibilities (e.g. tracking systems and student records systems). Plus there is an interest in using theory-based evaluation and case-orientated methodologies.

It may not benefit continued efforts to improve the data and evidence base for identifying an impact if the strongest types of approaches are conceptualised in a purely experimental sense, without consideration of alternative interpretations, because experimentation in the strictest sense is seen as risky and only relevant for access and participation work in some exceptional cases.

6.1.3 Evolving context for impact evaluation

Multiple factors were identified in the research as current considerations for access and participation evaluations and these have implications for evaluation designs. There are implications for the standard in terms of: i) constraints and opportunities for certain types of evaluation; and ii) the kind of impact evidence that providers are looking for to identify impact and to underpin evidence-informed decisions.

Current trends include:

- *Longer reporting timescales*
  The APPs allow for a longer reporting timescales which is an opportunity to engage in more longitudinal evaluations (especially as some outcomes take some time to show). The Type 3 standard relies on controlling for extraneous variables which is more challenging in complex social settings and when the impact is farther away in time from the intervention. The programme may have more control over short term outcomes, but to be impactful the relationship with the long term outcomes needs to be assessed.
- *Increasing concern for student involvement and OfS requirement for student views on APP*
  The standards focus on outputs and impacts which are defined in advance and turned into indicators and measures of impact, with an assumption that the goals chosen are 'right'. This interpretation is being challenged. Qualitative research is increasingly important to higher education delivery to ensure students have a positive experience. Co-creation activities are becoming more prominent to enable meaningful student voice with a concern to surface and learn from students' lived experience in order to understand what counts as effective.
- *Organisational/structural changes and the Whole Provider Approach*
  APP activities are becoming more embedded in the day to day activities of higher education. Intervention strategies may include interventions designed to bring about structural changes at the organisational level. Organisational changes may be difficult to capture in a quantitative sense. There are relationships between APP interventions and other strategies taking place at different levels: concerns for effectiveness come from different audiences and the impacts of specific interventions become harder to isolate.
- *Increasing provider diversity*
  Size of cohort makes a key difference to the types of data that are available on which to make conclusions. The delivery context needs to be taken into account as well, for example, whether coordination of evidence from different teams, sites/campus will be part of the picture. This can set up the conditions for comparative evaluation but also makes controlling the delivery more of an issue, and points to the need for contextualised approaches to evaluating.
- *Increasing focus on structural outcomes and impacts*
  The types of interventions go beyond those aiming for direct effect on a cohort of students/participants. The strategies encompass outcomes for staff and organisations, the outcomes of which can be harder to capture in a quantitative sense. APP Intervention Strategies draw together ways of tackling the same problem from different directions in order to maximise the overall impact. Interventions increasingly need to be looked at collectively.

The standards are based on a cause-and-effect approach to evaluating impact. They assume that impact can be attributed in a direct sense whereas intervention strategies are increasingly complex and embedded and part of the contribution to an overall goal (e.g. eliminating gaps). This becomes increasingly problematic in the current climate where providers are grappling with how to: deal with evaluation of complex programmes; get clarity on the purpose of impact evaluation knowledge development; link up the short, medium and long term impacts; use methods appropriate to the context; understand the best way to deal with the contextual factors in an evaluation; embed evaluation across whole systems; and involve students appropriately in evaluations.

There is a danger that the standards of evidence will become unfit for purpose unless they can support providers in the collection and use of evidence in the evolving higher education context. There are key questions to be addressed regarding: what counts as impact (contribution as well as attribution); the types of impact evidence that are most useful to support institutional decision-making; and how impact evidence is used alongside other types of evidence as part of cycles of reflection and review. Overall, there was a consensus that the standards will need to adapt to retain their role in transparency of decision-making.

The standards of evidence are based around the kind of claims that providers can make when backing up assertions of impact. The methodologies captured in different types of claims should be appropriate to all types of contexts and access and participation work (in order to have general relevance rather than amongst those looking to do certain evaluation designs). More guidance is needed on how alternative methods of evidence generation which may be required for complex types of access and participation interventions and contexts sit within the framework and what judgement can be made in terms of impact claims.

6.1.4 Use of evidence for external as well as internal knowledge generation

An expectation of evaluation for external publication is set in the APPs and this is blurring the lines between evaluation and research. Within institutions, there are advantages in conceptualising some types of ‘flagship’ evaluations which are designed to inform the external (as well as the internal audience) as research since it could encourage increased scrutiny, and more resources to undertake the evaluation and make better considered evaluation design choices. However, the amount of time/effort needed for external dissemination activities is unclear and there is a concern to make sure it is proportionate within an overall strategy so that evaluation most effectively supports delivery and outcomes for students.

In terms of assessing evaluation reports there is a difference between the use of standards as applied as an evaluation planning and development tool for APP evaluation work, and as a tool for retrospectively assessing the claims being made from specific impact evaluation reports. Using standards as a planning and development tool before/during an evaluation guides how the evaluation is conducted in order to ensure learning and improvement. Using standards as a retrospective assessment tool judges the validity of the causal evidence that has been produced in order to decide if an intervention is to be recommended more widely or not. If the distinction between these uses isn’t clear, there’s a risk that evaluation is judged unfairly (for example, criticising an approach for not meeting criteria it was never guided by). There are advantages to repositories of reports (for example the HEEL) categorising evaluations to help direct readers to how the knowledge can be used, but any assessments need to value the contribution from different evaluation contexts, and from different approaches to evaluation. Overall, there is perhaps an appetite for sharing wide-ranging evaluation studies cutting across all types of evidence and applying judgement based on their own merit or with reference to methodological quality criteria.

Decision-makers tend to lean towards evidence that reflects their particular circumstances and concerns. Useful evidence to support replication and transferability of interventions that have impact and associated effective practices needs to be contextualised to draw implications for others – e.g. particular target groups and delivery circumstances. If the knowledge is to help with replication, then understanding of the implementation and process factors needs to be given, alongside the evidence of impact.

6.2 Recommendations to OfS

In order to maximise the role of the standards of evidence as an evaluation capability building tool, we suggest there is scope for reinforcing and enhancing some parts of the existing guidance that might not always be recognised in how they are being applied.

Specific aspects of the standards requiring reinforcement have been identified as:

1) **Encouraging a cumulative approach:** The Types of evidence should be useful across the project planning and evaluation cycle when designed as cumulative and part of a layered approach to evaluating interventions. The layered approach involves:

Type 1: Ensure that there is a rationale for the intervention, and that there is understanding of what the outcomes and intended impact are, which is supported by evidence.

Type 2: Collect evidence to test whether the outcomes and impact from the intervention are better than might reasonably have been expected without the intervention.

Type 3: Collect evidence to show that it was the intervention that led to the outcomes and impacts rather than other factors.

2) **Complementarity with other types of evidence:** Evaluation of impact complements other types of evidence gathering that goes on to inform evidence-based decision making (such as process evaluation and practitioner reflection and so on).

In order to further support the use of standards in evaluation planning and delivery, we suggest there is scope to enhance the standards in these areas.

3) **Language of the standards:** The most useful evidence to support accountability and decision making should be at the forefront of the standards (rather than the use of any particular method or approach). We propose that the language of the standards should be revisited to include a strong focus on the claims that can be made - and the implications for how providers view the results from evaluation and to inform evidence-based decisions.

4) **Range of evaluation designs:** Within Type 2 and 3 it greater clarity is required to demonstrate that claims can be supported by a range of evaluation designs, recognising that qualitative approaches and case-orientated evaluation designs can be used to provide sufficiently strong evidence of impact claims in circumstances where the data and methods are sufficiently robust to do so.

Suggestions for change to the existing standards framework are given in Annex 2.

In order to encourage innovative approaches and methodological pluralism, as well as valuing contribution as well as impact, we suggest there is a need for:

5) **Greater clarity on methods to support claims to causality with the evidence types:** The standards typology is not meant as a categorisation of different methods of evaluation, but there is a lack of understanding about how different evaluation techniques can be used to support claims at all levels of the framework. The standards encompass quantitative and qualitative research methods within each of the types, but more information and guidance is needed on how different approaches can be operationalised in practice and how theory-based evaluation fits with the framework.

Recommendations for a framework to support these recommendations are given in Annex 3.

There is scope for further guidance to providers in relation to supporting evaluation strengthening and the use of evaluation evidence which could include:

6) **Practical guidance** on how providers can strengthen their impact evaluations within the Types (i.e. the practical ways to increase the rigour and quality of their evaluation design and methods). Those consulted in case institutions wanted guidance on how to operationalise the standards in different practice contexts.

7) The **evaluation self-assessment tool** should be revisited and revised in order to make sure it reflects a range of evidence that can be used (quantitative and qualitative) and that it encapsulates the most recent learning about the facilitating factors that can support strengthening of impact evaluations. The Tool could be supported by development of an **Evaluation Maturity Framewor**k which higher education providers can use to assess their capabilities and processes in APP evaluation to determine the level of advancement.

6.3 Recommendations to support the use of evidence

In order to encourage use of evaluation results in a way that supports transferability and replicability of proven and promising practices we recommend:

8) Differentiation between standards to make an assessment about the type of evidence **ex ante** (i.e. decisions about which evidence is appropriate in any particular context when planning and prioritising evaluation work) and the hierarchy by which to judge the strength of evidence **post hoc**

(i.e. the strength of causal claims reported for particular types of interventions in evaluation outputs). Ex ante standards help guide decisions and allocate resources wisely, while ex post standards ensure transparency, learning, and continuous improvement. The sector should be striving for the highest level of assurance on impact for both purposes, however, every evaluation needs to be judged on its own merit because every evaluation is an opportunity for knowledge generation and learning.

9) Clarity on expectations in terms of **contextual understanding** to be applied to impact evaluations. Reporting standards should keep take-up and transferability of knowledge in mind. A '+' designation could be included in the standards to denote evaluations where the contextual understanding and knowledge of practical replicability conditions is sufficient to underpin transferability across institutions (See suggestions in Annex 1).

10) Adoption of **reporting standards** designed to support take-up and transferability of knowledge. Reporting standards for published evaluations should help reviewers and readers understand and appraise an evaluation, taking account of the usefulness of the evidence as knowledge to inform others in the sector about what should happen next. There are potentially several aspects which need to be assessed, for example: Whether the evaluation questions have been clearly stated and are sufficiently defined; Whether the types of claims that the evaluation is seeking to make have been made explicit and that the evidence presented supports the claims; Whether the evidence and methods are reasonable, steps have been taken to ensure the reliability of the results, the evidence support the claims being made; Whether there is enough information about the contextual factors for providers to access the usefulness of the conclusions and transferability to their own practices; Whether there is enough information on the implementation aspects and processes involved to support replicability of the practice (if it has been shown to be worthwhile).

11) Reports of **evidence-based decision making practice** should be within the scope of the HEEL (i.e. studies which demonstrate where a provider used impact evaluation results to change their own practice and what was involved in that process including the evidence). This opens the door to learning about what different types of impact evidence studies have contributed in practice improvement and the implications of this learning for other providers.

6.4 Recommendations for further research

Further research would be beneficial, particularly in relation to:

12) Collaborative sector level research aimed at developing sector consensus on the specific requirements that different types of mixed methods and causal reasoning evaluation designs would need to meet to fit within thresholds within the standard types.

13) Commissioned research to identify learning from other evaluation frameworks. Potentially useful reference points for wider frameworks for complex evaluations are available from other sectors (e.g. the Medical Research Council (MRC)/ National Institute for Health Research (NIHR) framework for the development and evaluation of complex interventions in healthcare).

14) Developmental collaborative research to undertake further testing and validation of survey questions in existing evaluation frameworks in order to increase the availability of standardised tools for evaluating key sector intermediate student outcomes associated with access and participation interventions.

ANNEX 1: RESEARCH METHOD

The research was undertaken in partnership with a group of seven higher education providers which cut across different provider categories, contexts, and student populations (Table A1.1). The sample was purposively chosen to ensure a varied range of contexts and viewpoints across the English higher education sector. Evaluation leads in institutions were invited to participate as project partners at the end of 2024, through a direct approach from NERUPI (the Network for Researching University Participation Initiatives).

Table A1.1: Profile of Partner institutions

Student Group*	Finance Group*	Size of student body	Low Participation Neighbourhood (LPN) (*benchmark*)	Designation/ Mission group
High tariff	QI £100m-£200m	15,001–25,000 students	6.5 (*8.2*)	Research-intensive
Medium tariff	QI £100m-£200m	More than 25,000 students	6.2 (*5.9*)	Post-92/Million+
Medium tariff	QI over £200m and less than 70% of income	15,001–25,000 students	8.9 (*8.2*)	Research-intensive
Low tariff	QI over £200m and over 70% of income	More than 25,000 students	7.2 (*10.0*)	Post-92/University Alliance
Specialist: creative	Specialist: creative	Fewer than 5,000 students	13.3 (*15.0*)	Conservatoire
Specialist: other	Specialist: other	15,001–25,000 students	15 (*14.2*)	Private provider
High tariff	QI over £200m and less than 70% of income	10,001–15,000 students	7 (3.9)	Research-intensive Russell Group

*OfS categorisation

In the first stage of the research the partners took part in an initiation meeting in January 2025, at which the project was discussed. Colleagues began to share details of their approach to evaluating impact of access and participation activities. There were group work activities to consider how data and evidence were being used within the case institutions, and the implications for the standards of evidence and the project. The second stage involved a series of in-depth interviews with colleagues in partner institutions. Nineteen interviewees (mainly in-person) were completed in total broken down by role as follows: evaluation leads and evaluators (12); managers/leaders (including service and academic leads) (7). Interviews were designed as 'key informant interviews' – i.e. targeting colleagues recognised for their insider knowledge and unique perspectives on the topic. This method is distinct in focusing on information-rich sources and aiming for depth of insight rather than breadth. A semi-structured interview script was used which included both general questions and provider specific questions. The interviews were supported by desk research to draw further insights into the evaluation approaches within the institutions including scrutiny of the latest APP documents.

The fieldwork topics were wide-ranging but included a concern to ascertain:

1. What effects are the standards of evidence having on current approaches to evaluation?
2. How do decision-makers in universities obtain knowledge about effective practices, what information do they need, and what role does impact evaluation play in this?
3. How are the standards supporting evidence-based decision making within institutions?
4. How are the standards supporting understanding of replicability and transfer of proven and promising practice including knowledge transfer across different institutional contexts?

The data from the interviews and desk research was analysed in two ways:

1. Qualitative analysis working towards exploratory findings with the aims of extrapolating some conclusions (and making some tentative generalisations).
2. Identification of specific approaches and methods in examples of evaluations in order to explore decisions and approaches in different settings where different issues and solutions are experienced (in order to describe and explain approaches to undertaking impact evaluations and the use of evidence in decision making).

The research partners were invited to comment on draft materials and the group met twice to discuss and agree the findings and recommendations. The research reporting and development of conclusions and recommendations was supported by a reference group of evaluation experts, senior institutional leaders and representatives of sector networks (HEAT). This involved a series of three meetings and exchange of materials for comments and agreement.

ANNEX 2: SUGGESTED ENHANCEMENTS TO THE STANDARDS FRAMEWORK

	Criteria to meet this Type	Claims that can be made[15]	Evidence required	
Type 1:	**Explains what is being done and why** The impact evaluation provides a narrative or a coherent theory of change to motivate the selection of activities in the context of a coherent strategy	**Claims of policy** There is a coherent explanation of what we do and why based on research and reasoning	**1a) Evidence of Impact Elsewhere or in Research Literature** You have evidence of impact elsewhere and/or in the research literature on access and participation activity effectiveness **1b) Proven or Promising Practice from Existing Evaluation Results** You've run the intervention before and have internal evaluation data showing it worked—or at least showed promise. **1c) Logical Causal Chain with Corroborating Evidence** You have defined a logical causal chain for outcome/impact and have evidence to corroborate that the intervention could be expected to bring about a positive change. **1d) Emerging Evidence of Beneficial Results** You have initial evaluation results showing that the activities are related to beneficial results in line with the objectives.	Reasoned Practice
Type 2:	**Shows positive outcomes without proving causality** The impact evaluation collects data on impact and reports evidence that those receiving an intervention have better outcomes than might otherwise be expected	**Claims of worth** We can demonstrate that our intervention is associated with beneficial results against a counterfactual.	**2a) Pre/Post Change or comparison to Non-Participants** You have quantitative and/or qualitative evidence of a pre/post intervention change or a difference compared to what might otherwise have been expected. **2b) Causal Reasoning with Programme Theory** You've defined a logical causal chain for outcome and impact and systematic quantitative and/or qualitative evidence from causal reasoning which demonstrates that implementation of the programme theory for your intervention plausibly explains the outcomes in your context*	Promising Practice

[15] The focus of the standards is on setting on the evidence the providers would need to back up their claims to impact. However, if an evaluation generated evidence of no (or negative) impact there will still be important learning about what (doesn't) work.

Type 3:	**Demonstrates that the intervention caused the observed outcomes** The impact evaluation methodology provides evidence of a causal effect of an intervention	**Causal claims** We can demonstrate our intervention causes improvement and the difference it makes	**3a) Quasi-Experimental Design with an Appropriate Comparison Group** You have a positive result from of a treatment change on participles relative to an appropriate comparison group who did not take part in the intervention from a quasi-experimental design** **3b) Randomised Controlled Trial (RCT)** You have a positive result from of a treatment change on participles relative to a control group who did not take part in the intervention from a randomised experiment. **3c) Non-experimental Causal Inference Methods** You have a defined logical causal chain for outcome/impact and have systematic quantitative and/or qualitative evidence to demonstrate that the intervention is an explanatory variable for the observed outcomes in your context and you have ruled out alternative explanations.	Validated Practice

* Theory-based evaluation designs included in this type of evidence (see discussion in Annex 2).
**Within a quasi-experimental design whilst causation may not be proven (because there may be other differences between the groups that caused the effect) this is still a strong design if an RCT is not appropriate: QEDs have the potential to point in the direction of more generalisable results so long as the contextual factors are understood, and are usually less costly and time-consuming than an RCT.

'+' Designation to support transferability of practice where there are proven/promising results

In order to support the use of evaluation to inform decisions on access and participation activities across the sector, we suggest that the evidence base for different types of evaluations can be further enhanced to achieve a '+' designation (e.g. Type 3+).

	Criteria to meet this Type	Claims that can be made	Evidence	
'+' designation	The impact evaluation is underpinned be detailed articulation of the considerations required to faithfully replicate it.	**Claims of transferability** We understand enough about effectiveness in context and the systems and procedures to ensure consistent replication and positive impact	You can explain how and why the intervention had the observed impact on the access and participation objectives it was designed to address and can demonstrate the requirements in terms of implementation and delivery that are needed for the positive impact to be achieved in practice.	Replicable Practice

*Note transferability requires understanding of both the original intervention context and the new setting in order to consider whether an intervention that worked in one context can be adapted and still be effective in a different context.

ANNEX 3: INTERATION OF THEORY-BASED EVALUATION FOR COMPLEX INTERVENTIONS

According to the Magenta Book, theory-based methods are appropriate when experimental or quasi-experimental designs aren't feasible, and they can still meet high standards of evidence if they: clearly articulate a theory of change; gather robust evidence to test causal mechanisms; rule out rival explanations.[16] Existing theory of change guidance already puts a focus on theory-based evaluation but clarity is needed on how to evaluate the impact of a programme design using a theory of change approach.

Suggestions for how the standards of evidence could be achieved using 'hybrid evaluation methods that combine outcome/impact evidence with theory-based evaluation methods are given here.

'Hybrid' evaluation methods

	Type 1	Type 2	Type 3
Standards of evidence focus	How does the programme make a difference to the outcomes?	The difference compared to a counterfactual estimate?	The difference that can be attributed directly to the programme?
Implementation and Process Evaluation (IPE) focus*	Did the programme reach the target group(s) and how was it implemented and experienced in practice		
Theory-based evaluation focus	How was the programme implemented? Were there deviations from the original plan? (Formative)	How does the implementation affect the outcomes? What does good implementation look like? (Summative)	Understanding what was implemented and what are the implications in terms of the causal effect of what was implemented on the outcomes in context? (Summative)
Examples of Hybrid methods?***	Literature Review to articulate the rationale and link it to existing evidence of effectiveness. Theory of Change + stakeholder input to co-create/justify the intervention logic. Contribution Mapping to show how the programme fits within a wider system of change.	Theory of Change + Pre/Post Survey (measure change in confidence or engagement before and after participation) Realist Evaluation using mixed methods (explore what works, for whom, and in what contexts). Process Tracing (quant & qual evidence to test steps in the theory of change)	Realist Evaluation + Difference-in-Differences (explore causal pathways while comparing outcomes over time between groups). Process Tracing + Matched Comparison (mechanism testing plus quantitative outcome analysis). Theory of Change + Instrumental Variables (to isolate causal impact).

* Process evaluation doesn't estimate causal effects in the statistical sense, but can: validate causal mechanisms; confirm that the steps in your theory of change actually occurred; explore contextual factors; understand why and how the intervention worked (or didn't) for different student groups; identify unintended consequences; surface outcomes not captured by quantitative outcome indicators; strengthen causal contribution claims:

** Type 3 evaluation requires evidence of causal impact, not just association. Theory-based methods can help rule out alternative explanations and reinforce the plausibility of impact when experimentation is unsuitable.

*** Qualitative insight uncovers mechanisms, conditions, and context. Quantitative evidence isolates impact and generalizes findings. Together, they produce a narrative that is methodologically rigorous and meaningfully rich.

[16] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879438/HMT_MagentaBook.pdf