Feasibility of enhancements to the Office for Students (OfS) Standards of Evaluation Evidence for Access and Participation Plans

Executive Summary

November 2025

Introduction

The Office for Students (OfS) standards of evidence set out a framework for assessing evidence about the impact of widening participation policies, practices and programmes[1]. Higher education providers agree Access and Participation Plans with the OfS detailing the strategies and activities that they will implement over a four-year period to address barriers to access and success, along with how they will meet reporting requirements for closer monitoring and dissemination of equality, access and participation evaluation results, with reference to the standards of evidence. The Higher Education Evaluation Library (HEEL) is being developed as a repository for sharing evidence.

In this context, the aims of the research were to:

1. Inform national OfS guidance on evaluation methods supporting transfer of best practice from impactful programmes;
2. Enhance universities' capacity to produce effective evaluation evidence for widening participation interventions and fulfil their APP commitments;
3. Develop recommendations and tools enabling the transfer of effective practices across the higher education sector.

This was a collaborative project funded by the UKRI Policy Support Fund (PSF) 2024-25. It involved 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 were involved in agreeing the outputs.

The research demonstrates that the OfS standards of evidence have played a key role in building evaluation capability across the English higher education sector. Changes since the standards were developed, especially their application across the student lifecycle, coupled with the complexity of applying causal designs in complex education environments, suggests review of the standards is needed. There is scope for reinforcement and enhancement of the standards, alongside more practical guidance on evaluation strengthening, further evolution of tools to support evaluation capability building, and developments to clarify expectations on reporting evaluations.

The standards of evidence framework defines three types of impact evaluation evidence:

Type 1) Narrative - a coherent explanation of why the intervention is expected to work and why;

Type 2) Empirical - data to show those receiving an intervention have better outcomes;

Type 3) Causal - provides evidence of a causal effect.

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

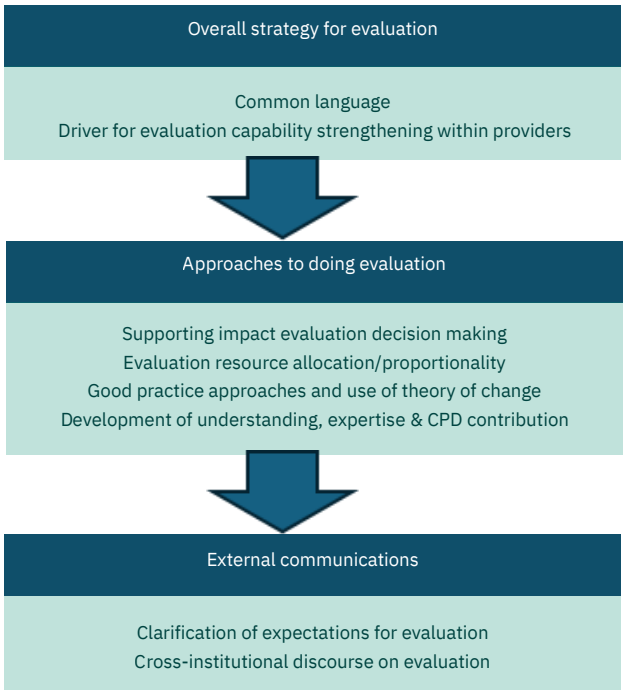
Key findings: Implications of the standards of evidence

The standards of evidence have supported evaluation capability-building and communication on evaluation practice within higher education providers to enhance how they review their access and participation work. The standards have underpinned development of the overall evaluation strategy; informed decisions to evaluate specific interventions and the types of evaluation that are put in place; and have supported enhanced understanding of what counts as useful evidence for decision-making amongst managers, practitioners and evaluators.

The standards of evidence and the associated self-assessment tool have pushed providers to implement increasingly robust approaches to assessing their interventions' contributions to outcomes, providing greater surety that interventions are making a difference. The standards inform decision-making at project, programme and institutional levels. They have supported planning and use of good practice approaches to building the rationale for interventions and clarifying the outcomes and impacts using theory of change.

There are tensions for evaluation planning on the ground: on the one hand there is a need to strive for strong evidence across the field (at least Type 2 for intervention strategies as a whole). On the other hand, there is a clear need to prioritise (limited) evaluation resources and ensure that evidence is used to the greatest effect. Evaluating all activities to the same extent continuously is disproportionate (although successively strong evidence could emerge over time). Uncertainty existed around the expectation to evaluate interventions at the project/programme level or by setting in place an overarching approach for the entire intervention strategy (or both).

Figure 1: Benefits of the standards of evidence



The strength of evidence required for different types of decisions is a consideration, recognising that more reliable evidence of impact will be required for the important decisions (e.g. whether to commit significant resource to an intervention or completely change what is delivered). Type 1 evidence might be used to end an intervention as there would be little benefit seeking to identify differences in outcomes if initial data suggested only poor outcomes were being achieved. However, other types of causal evidence would be useful when deciding whether to continue one intervention over another. More surety might be required that the intervention caused the impact

Figure 2: 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 what could have been expected 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
Evidence type(s)	**Type 1**	**Type 2**	**Type 3**	**Type 1/2**

if the institution was considering rolling-out a pilot or recommending the practice to other providers. Evidence below direct causal inference might be considered sufficient if the method was rigorous and proportionate. For example, counterfactual evidence could provide significant insights alongside other forms of reliable evidence of success, and/or convincing evidence that the programme theory worked.

Impact evaluation complements other types of evaluation and is used in a formative and summative sense as part of ongoing cycles of reflection and review. Evidence that is layered to 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.

The standards act as a bridge between internal learning and external accountability. They help programme managers, decision-makers and staff to reflect on delivery and make improvements whilst ensuring that the internal processes meet sector expectations, foster accountability for the investments, and help make findings credible to outsiders. Interviewees said the standards provide a 'common language' and a 'touchstone' for the sector on evaluation, however, there is a 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 risk that the standards are a 'shorthand' for a methodological distinction between experimental, quantitative and qualitative methods and are applied not just to impact evaluation but to evaluation in general.

Key findings: Types/Levels Discourse

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 access and participation evaluations. Experimental methods are a strong design when it comes to demonstrating causality but all types of evidence provide very valuable knowledge to inform decision making on APP interventions in practice and there is 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.

The case study research highlighted the tensions in the current standards about the optimum way to move to stronger causal surety. The most expensive interventions warrant the strongest type of evidence, but tend to be complex and beyond the scope of RCTs which are perhaps best applied to discrete activities with direct outcomes. Experimentation is seen as problematic and only relevant for access and participation work in some exceptional cases. Ethical considerations, practical issues, and data constraints were identified as arguments against RCTs. There is a danger of focusing on 'measurables' above other meaningful outcomes, and the high possibility of an inconclusive result makes experimentation risky. Quasi-experimental approaches are supported by developments in data possibilities (e.g. drawing on tracking systems and student records systems). Moreover, many APP interventions lend themselves towards case-orientated methodologies and theory-based evaluation - i.e. evaluation which aims for systematic understanding of how and why an intervention works, focusing on the causal mechanisms and underlying assumptions, as well as the expected outcomes. A 'layered' approach to impact evaluation integrates causal discovery, causal reasoning and causal inference methods to test the programme theory underpinning interventions (Figure 3).

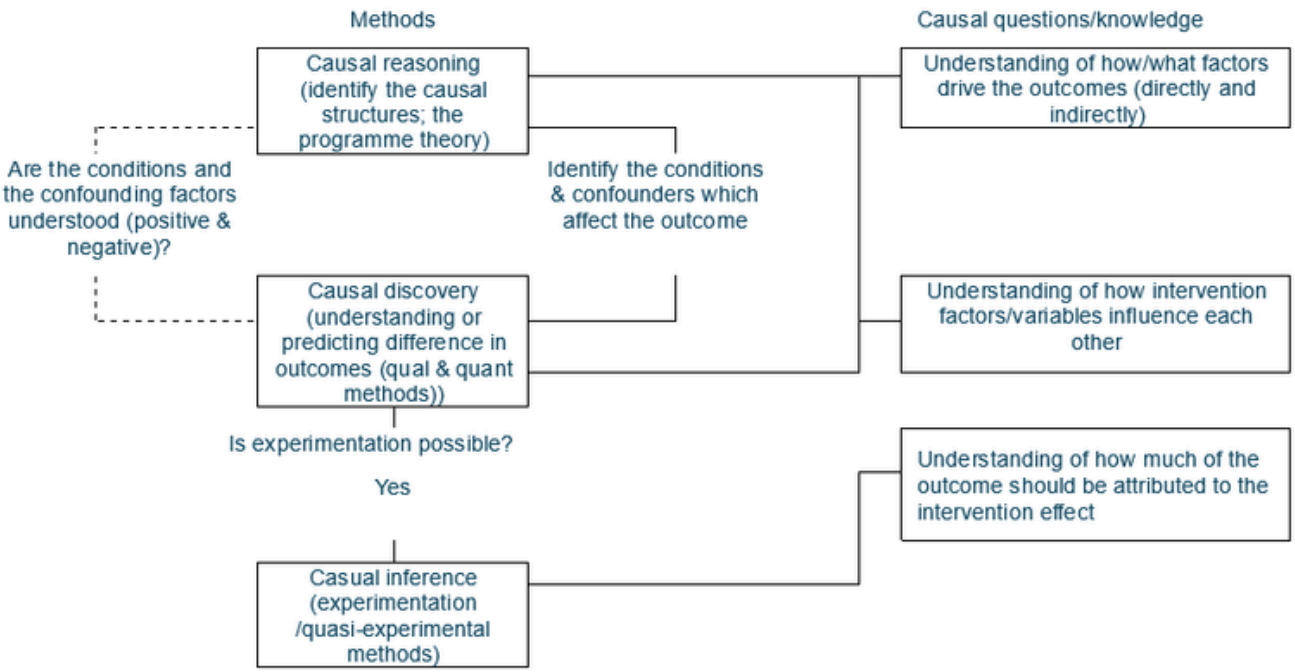
Recommendations to the OfS; Reinforcement of the Standards

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 implementation and process evaluation (IPE), practitioner reflection and so on).

Figure 3: Causal Questions and Evaluation Methods



The standards of evidence specify the claims that providers can make when backing up assertions of impact. 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. The methodologies captured in different types of claims should be appropriate to all contexts in order to have general relevance rather than amongst those looking to use experimental evaluation designs.

Guidance is needed on how alternative methods of evidence generation required for complex types of access and participation interventions sit within the standards framework and what judgement can be made in terms of impact claims. Within Type 3 in particular it is unclear how non-experimental causal inference methods (i.e. the different methods that can be used to identify the causal structure between variables which do not manipulate the study environment) should be viewed and whether causal inference with qualitative data can meet the standard.

Recommendations to the OfS; Reinforcement of the Standards

3) **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 should view the results from evaluation to inform evidence-based decisions.
4) **Greater clarity is required** to demonstrate that claims can be supported by a range of evaluation designs, including qualitative approaches and case-orientated designs if they provide sufficiently strong evidence of impact claims in circumstances where the data and methods are sufficiently robust to do so.
5) Suggestions for revisions to the standards framework are given in **Box 1**.

   Collaborative sector level engagement is needed to develop sector consensus on the requirements that different types of mixed methods and causal reasoning evaluation designs need to meet to fit within thresholds within the standard types.

   Suggestions for 'hybrid' methods are given in **Box 2.**

   Learning can be identified from evaluation frameworks in other sectors. For example, a useful reference for complex evaluations with transferable learning is the Medical Research Council (MRC)/National Institute for Health Research (NIHR) framework for the development and evaluation of complex interventions in healthcare).

Box 1: Enhancements to the standards framework

	Description	Type of claims*	Evidence required	
Type 1:	**Explains what is being done and why** Impact evaluation provides a narrative or 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 you 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 can be expected to generate positive change. **1d) Emerging Evidence of Beneficial Results** You have initial evaluation results showing the 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** You can demonstrate that your 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 have 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
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** You can demonstrate your 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 participants 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 participants 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 defined a 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

*The standards categorise the evidence the providers need to back up their claims to impact. If an evaluation generated evidence of no or negative impact then the standard would not have been met, even though there will still be important learning about what (doesn't) work.

**Theory driven designs included in this type of evidence (see discussion in Box 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.

Box 2: Use of 'hybrid' evaluation designs

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[2]. Existing OfS guidance already puts a focus on use of theory of change but clarity is needed on how to evaluate the impact of a programme design using a theory-based evaluation approach.

Some 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.

	Type 1	Type 2	Type 3
Standards of evidence focus	How does the programme make a difference to the outcomes?	What's the difference compared to a counterfactual estimate?	To what extent can the difference be attributed directly to the programme?
Implementation and Process Evaluation (IPE) focus*	Did the programme reach the target group(s) and how was it 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)	How can we understand the causal effects of what was implemented on the outcomes in our context in order to identify whether the theory of change played out as expected? (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 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. Other 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.

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

Key findings: Opportunities for evaluation strengthening

Evaluation is influenced by the context of the higher education provision. However, common themes emerged from case institutions:

Evaluation design challenges: struggles to prove causality in complex educational environments; difficulty measuring intangible outcomes; limited evaluation frameworks especially those tailored to creative, vocational or specialist settings; and limitations in expertise and resource which affected evaluation design possibilities.

Data analysis and interpretation issues: difficulties analysing and making sense of outcomes data when there is no access to a counterfactual or comparison group, controlling for extraneous confounding variable, understanding inter-sectionalities/complex student demographics. Controlling for the implementation factors were particularly challenging in large institutions with devolved decision-making.

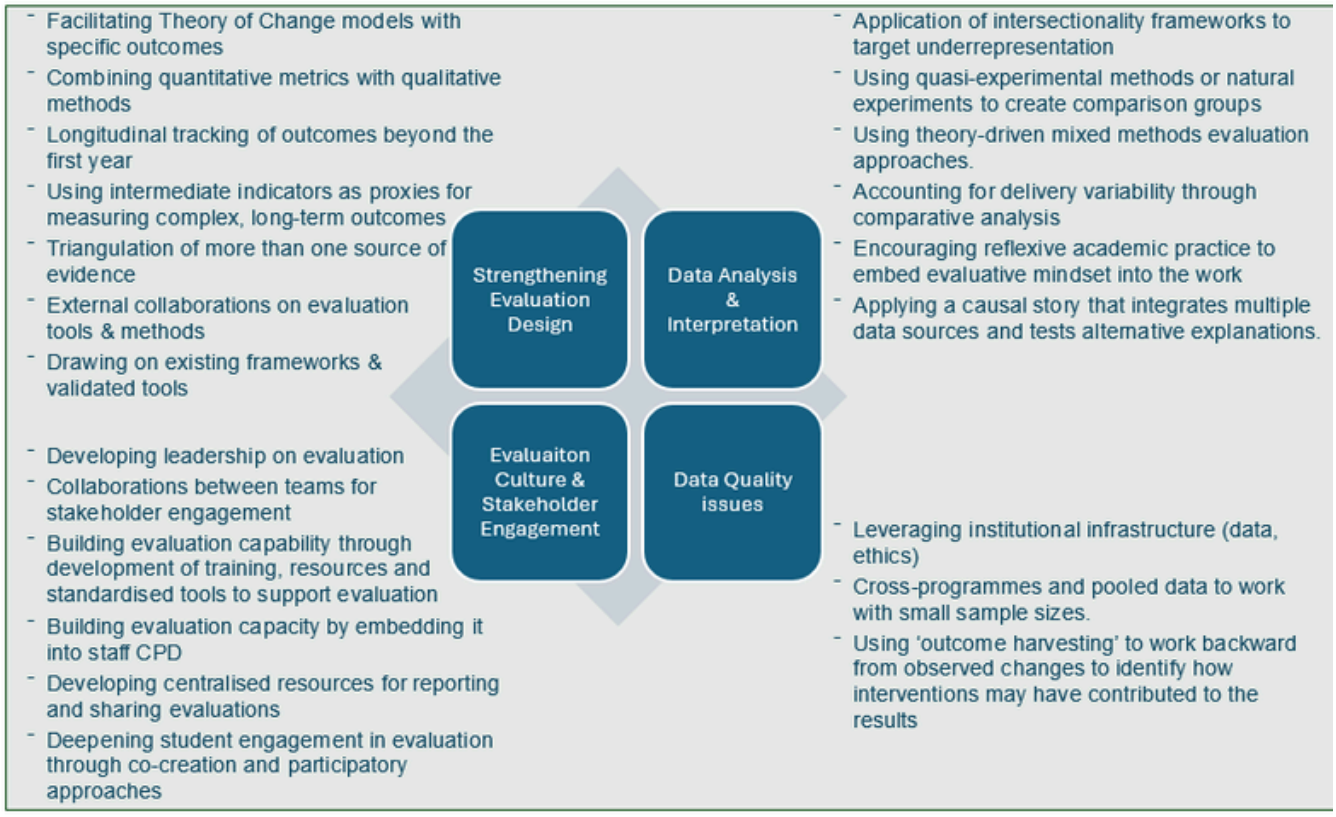
Data quality issues: challenges with accessing high-quality granular data on student outcomes; difficulty tracking long-term; non-linear progression, and designing scalable evaluations with small sample sizes.

Evaluation culture and stakeholder engagement challenges: lack of institutional leadership on evaluation; patchy engagement from academic staff and practitioners; reliance of self-reported measures; inconsistencies and difficulties in engaging students in evaluations, and the need to work through gatekeepers.

All the case institutions were working to embed evaluation into the design and delivery of initiatives, rather than treating it as an afterthought. Evidence strengthening was going on in terms of evaluation capability building developments institutionally, and in approaches to evaluating particular projects and programmes. Helpful practices identified in the research to address the challenges and strengthen the evidence base are summarised in Figure 4.

Evaluation strengthening is being underpinned by a cultural change within case institutions, encouraging staff to see APP evaluation as a tool for learning, not just accountability. Policies and resources are needed to support cultural transformation and stakeholder engagement, and these take time to get in place. Embedding evaluation as part of institutional-wide practice requires strong leadership commitment to data collection and leverage of institutional processes and data systems.

Figure 4: Opportunities for evaluation strengthening



Undertaking evaluation purely for regulatory purposes has potentially limited value internally, adversely affecting engagement and capacity for effective evaluation. When used strategically to inform action, the value of evaluation is recognised and opportunities develop for wider engagement in evaluation emerge.

Clarity about the benefits of evaluation for different stakeholders is helpful: practitioners are looking to adjust delivery based on real-time insight; those in strategic roles are aiming to inform broader programme design and resource allocations; and governance teams want to ensure alignment with institutional and OfS priorities. Well-designed formative and summative evaluation as part of an iterative evaluation approach across the intervention lifecycle can meet the needs of different audiences – bridging the gap between internal learning and external accountability.

Suggestions for institutional leaders

Suggestions for evaluation strengthening emerging from the research included:

- Design for evaluation from the start: use theory of change in project planning, specify outcome measures and define the data points, focusing on priorities that matter and that help to identify the causal pathways by which the investment makes a difference.
- Measure what matters: use prioritisation tools to tailor the overall evaluation effort appropriately and match evaluation depth to intervention maturity. Report gaps and inter-sectionalities as standard.
- Grow evaluation capability through academic practice: embed APP evaluation expertise into staff professional development, link staff qualifications with opportunities for conducting APP research and evaluation, encourage practitioner-research projects, build internal Communities of Practice.
- Focus on utility to different (priority) audiences: use findings in real time, embed evidence into decision-making cycles throughout the project lifecycle—not just at the end.
- Blend qualitative and quantitative methods in order to develop a rich picture, to strengthen the generalisability of findings and explanatory power.
- Build expertise and infrastructure for evaluation: establish central evaluation support, use dashboards or databases to track interventions and outcomes, invest in shared tools to support data collection, linkage, tracking and analysis, foster cross-team collaboration and evidence sharing.
- Identify what most matters to students, include student narratives in evaluation.

Impact evaluation for external knowledge generation

The standards were designed to build knowledge on the most impactful activities in order to direct funds effectively. An expectation of evaluation for external publication is set out in the APPs. However, the research suggests that increased emphasis on evaluation for external as well as internal audiences blurs the lines between evaluation and research. Within institutions, there are advantages in conceptualising evaluation of some types of 'flagship' interventions as

Recommendations to the OfS: Guidance and tools to support evaluation strengthening

6) Provide **practical guidance** for providers to strengthen their impact evaluations within the Types (i.e. the ways to increase the rigour and quality of their evaluation design and methods and the pitfalls to avoid/mitigate).

7) **Revise the Evaluation Self-Assessment Tool** 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 Framework which higher education providers can use to assess their capabilities and processes in APP evaluation to determine the level of advancement.

8) Further encourage developmental **collaborative research** to test and validate of survey questions in existing evaluation frameworks supports evaluation strengthening by increasing the availability of standardised tools for evaluating key sector intermediate student outcomes associated with access and participation interventions.

sector knowledge-building research since it encourages greater scrutiny, levers evaluation resources and better considered design choices. Dissemination should be proportionate within an overall strategy in order to most effectively support delivery and outcomes for disadvantaged students within the institution.

There are advantages to reports in the HEEL being categorised, but any assessment should value the contribution from different contexts equally. There is a difference between using standards as a tool for retrospectively assessing claims to impact, and for evaluation planning and development. The latter guides how the evaluation is to be conducted and anticipates the impact before the evidence is available for scrutiny. The former informs if there is sufficient causal evidence to recommend an intervention (or not). If the distinction isn't clear, there is a risk that ex ante evaluation is judged unfairly (e.g. criticising an approach for not meeting criteria it was never guided by). Overall, there appears to be an appetite for all types of evidence valued on relevance as well as with reference to methodological criteria.

Decision-makers lean to evidence that reflects their circumstances and concerns. Evidence to support replication of impactful interventions and effective practices needs to be contextualised (e.g. with reference to target groups and delivery). For findings to be transferable, understanding of implementation and process factors is usually needed, as well as evidence of impact.

Recommendations: Guidance to support use of evidence

9) Standards to make an assessment about the type of evidence **ex ante** (i.e. which evidence is appropriate in when planning and prioritising evaluation work) should be differentiated from the hierarchy ex post to judge the strength of claims to impact (i.e. evidence reported in impact evaluation outputs). Ex ante standards help to guide evaluation decisions wisely, while ex post standards ensure transparency, learning, and continuous improvement.

10) **Evidence standards** for published evaluations should help readers appraise the strength of evidence, taking account it's usefulness as knowledge about what should happen next. Contextual understanding is needed in impact evaluation reports, keeping the take-up and transferability of the 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 of the conclusions (see **Box 3**).

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). 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.

Box 3: '+' Designation to support replication where there are proven/promising results

To support use of evaluation to inform sector decisions on APP activities, we suggest that the evidence base for different types of impact evidence can be enhanced to achieve a '+' designation (e.g. Type 3+).

	Criteria to meet this type	Claims that can be made*	Evidence required	
'+' designation	The impact evaluation is underpinned by detailed articulation of the considerations required to faithfully replicate it	Claims of transferability* You understand enough about effectiveness in context and the systems and procedures to ensure consistent transferability and scalability for 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	Transferable 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.

Authors
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

This work was supported by Research England under the Policy Support Fund (PSF) 2024-25 funding stream.