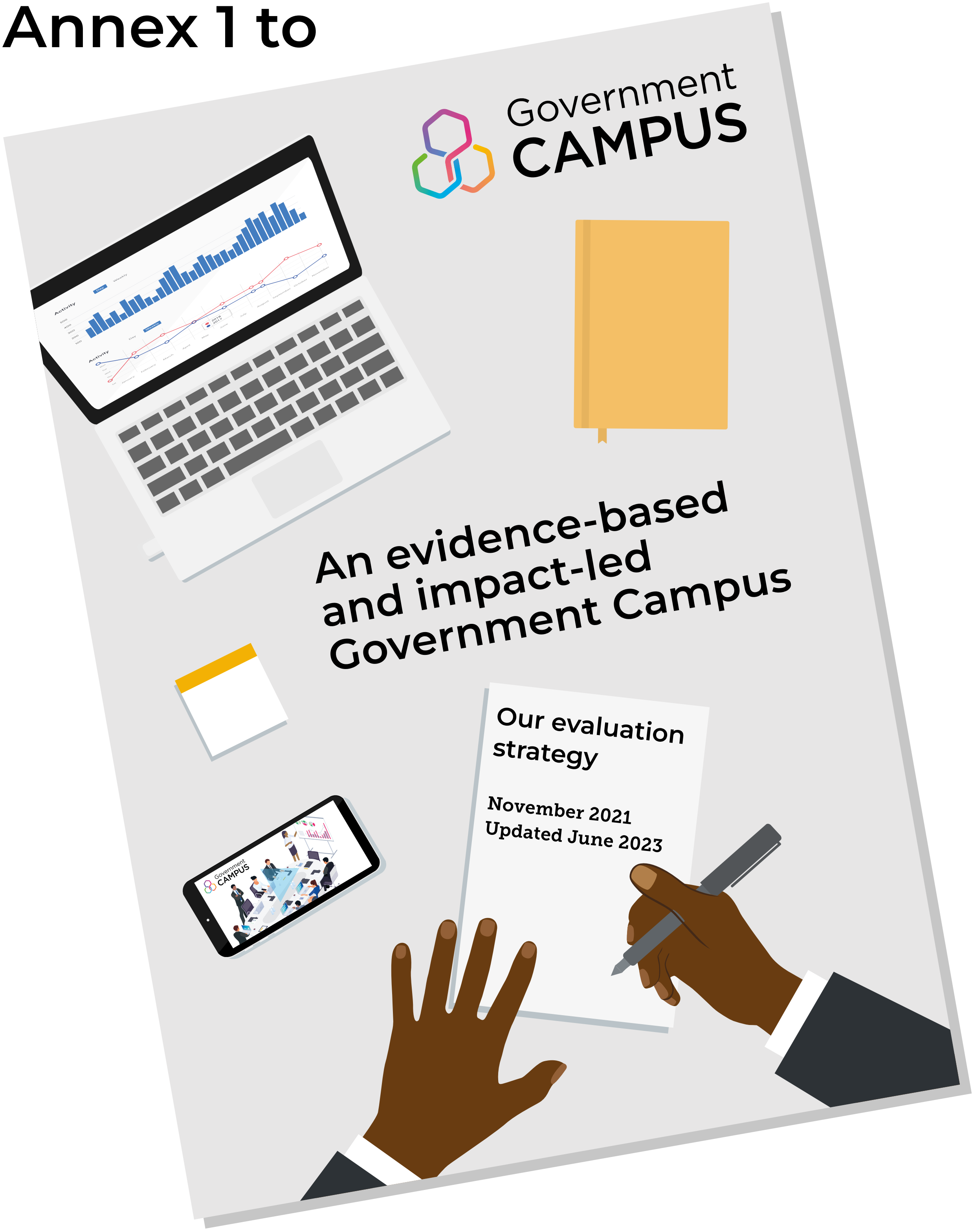
Government CAMPUS

Annex 1 to



Government CAMPUS

An evidence-based and impact-led Government Campus

Our evaluation strategy

November 2021
Updated June 2023

Design Principles for Evaluations

Last updated June 2023

Design Principles for Government Campus evaluations

General Approaches

1. All monitoring and evaluation activities should **follow current civil service policies**, including the Magenta Book. GSCU holds an index of relevant policies which is reviewed regularly. This is supplementary guidance.
2. The monitoring and evaluation needs of all GSCU interventions must be assessed by an Evidence and Impact specialist within GSCU (that is, someone in an analytical role, responsible for delivering our evaluation strategy, who is badged to a Government analytical profession), who will work with the programme team/s to determine what is required and how it is best delivered. The evidence and impact specialist will determine the level of delegation which is appropriate for leadership and management of the evaluation, based on its scale and complexity. We strongly recommend that evaluations of Campus products and programmes where GSCU is not the main commissioner/owner (i.e. products procured by departments/professions through the central Learning Framework contracts and other curriculum-linked activities owned by departments and professions) have a badged analytical professional acting as the internal civil service lead and coordinator for the evaluation.
3. As outlined in our evaluation strategy our overall approach to evaluating the Government Campus and its constituent programmes, activities and products is based on **deep subject matter expertise**. That is, research and evaluation methods and measures should be built out of deep knowledge and understanding of the academic evidence base and methodologies for research into learning, professional development, behavioural change and organisational change. Generic methods and measures should be adapted for the specific context, including by ensuring evaluation designs are based on systematic understanding of how these sorts of questions and outcomes have been evaluated before, in credible studies, within the fields of learning sciences, behavioural science etc.

4. Attention should be given to the likely **future trajectory** of the intervention (particularly, when decisions need to be taken on this and/or any similar programmes) and its **history** (why has it been designed in this way, what policies is it linked to, what has been tried and not/worked before), to **design an evaluation which is as useful as possible in generating evidence at the right time for future decisions, and asks the right questions**. The level of evaluation should **always be proportionate to the likely impact the evidence generated in the evaluation can have on influencing future strategic decisions and spend**. This is because the role of evaluation in government is not to generate 'pure' research which contributes to wider knowledge for its own sake, but to undertake 'applied' research which supports effective government. (Partnerships with academic institutions and external funders may be able to contribute to both these types of research in different ways.)

5. **Research before evaluation:** Programme (re)design should be based on review of robust evidence from prior evaluations, so that it is evidence-based, and thus has the greatest chance of being effective. Similarly, the focus for an evaluation should be informed by this external evidence. It is a waste of resources to use trials and specialised quasi-experimental methods to answer questions that can be answered using relatively inexpensive and unobtrusive desk based research of robust external evidence. The level and rigour of evidence synthesis should be proportionate to the scale and importance of the intervention and the level of innovation/evidence uncertainty. For example, if the design for a new programme is similar to an existing design, for which evidence has already been reviewed, it would not be necessary to review this evidence again, only update it. Field-specific experts (e.g. they work in research on leadership, coaching, professional learning etc.) will usually be best placed to synthesise the relevant literature and its features quickly and accurately. They may be internal or external.

6. **A theory-of-change driven approach** should be taken to programme design, to ensure that design is impact led and to support the planning of purposeful evaluation, i.e. expected outputs, outcomes and impact will be specified, and linked to the inputs/activities; and the assumed pathways to change, levers and barriers will be made explicit, and challenged with internal and external evidence (1, above).

7. **Evaluation should be planned from the outset of programme (re)design**, and sufficient time should be allowed for scoping, design, feasibility testing and (where appropriate) commissioning of an appropriate method. The first stage of evaluation scoping will be to assess which tier of evaluation is likely to be appropriate.

[1] Robust' is understood to mean that it uses only high quality external evidence (academic and robust 'grey literature'), prioritising well-conducted systematic reviews and meta-analyses, followed by other high quality systematised review methods, before turning, where necessary, to individual study results, which have been appraised for quality before inclusion. It should be undertaken by a qualified individual with relevant field-specific expertise, and should use a transparent, replicable, appropriate and systematised method for appraising and synthesising the literature.

All projects should start with **scoping of existing (quantitative and qualitative) data sources**, including 'naturally occurring' administrative data, and an identification of what could be accomplished with this existing data, before embarking on any new data collection. If operational reasons mean that evaluation was not able to be planned from the outset of a design process, Evidence and Impact Leads should be consulted as early as possible so they can support with embedding evidence and evaluation within any operational constraints.

8. **Monitoring and evaluation which makes use of existing administrative data and/or makes use of data collection which is smoothly embedded in programme delivery** are always preferred over new data collection, unless new data collection (e.g. participant surveys or interviews) is clearly required to address evaluation questions which cannot be addressed through other sources.

9. Careful attention should be given to **outcome measures**, particularly ensuring that the nature of an outcome has been carefully specified to a sufficient degree to reliably measure it, and that a suitable outcome measure has been identified which is valid for measuring that outcome. For example, broad statements like 'improved management', 'good communication skills' or 'systems leadership' are not measurable outcomes unless what they specifically look like in practice is specified. Pre-validated and pre-existing evidence-based measures should always be sought out, and would always be preferred to designing new measures. If new measures need to be designed expert support should be sought to ensure they are valid and reliable, including conceptually valid for the outcome in question. **Colleagues should try to use the same outcome measures across multiple evaluations which are interested in similar outcomes**, to allow for cumulation of findings.

10. In most circumstances, **impact and process evaluation should be undertaken in parallel**, unless there is a particular reason not to. **Process evaluation** should (especially in the upper tiers) go beyond basic implementation measures or fidelity questions and be focused on why a programme does/not work. For example, a well-targeted question about implementation, causal pathways or contextual variation, in line with our realist approach, ensuring the focus of the question and the method chosen will enable cumulation within our wider Campus evaluation programme. Whilst this requires additional resource, by adding depth it will ultimately make the findings more useful across the wider civil service context, and this is preferable to devoting resource to a greater number of 'shallower' and less generalisable or cumulative studies.

11. The level of **economic evaluation** required will be determined by the tier of evaluation, but in all cases colleagues should be clear on the 'input' cost, ensuring this includes all inputs (e.g. time and resources as well as design or delivery cost).

12. Campus **ethical governance processes** for research and evaluation must be followed from the outset and built into every stage of design and delivery, alongside the usual ethical and legal guidelines and protocols that govern social research in Government, published by GSR.

Design Principles for Tier 1 evaluations:

Once it has been established that a programme or area of activity will require a Tier 1 evaluation, the following principles should be followed:

1. The nature of Tier 1 evaluations is that the spend and strategic importance of the programme means that there is a necessity to account for the investment via robust impact and economic evaluation, as well as to learn lessons for the future through the impact and process evaluations. Tier 1 evaluations must use the most robust methods. Exactly what methods will be appropriate will be determined by the level of uncertainty in the evidence-base for the intervention and its applicability in our context. The methods are very likely to entail using an appropriate **experimental or quasi-experimental design**, as defined by the **Magenta Book**. Appropriate systematised scoping of credible external literature should be conducted to identify how this method has been successfully applied in the field of professional learning by other researchers (or could be successfully applied) before embarking on any evaluation. Where there is uncertainty about the application of the method to the intended context, the relevant analytical lead should consider commissioning a methodologically-focused scoping review to inform decision making.

2. At this tier we would generally expect the intervention to be based on robust and up to date synthesis of high quality evidence, for example in a Systematic Review, Meta-Analysis, Realist Synthesis etc. Where such syntheses already exist we would expect those to be used, rather than commissioning duplicative review work. Proposed evaluation foci should be clearly linked to the theory of change and areas of uncertainty based on the prior evidence.

3. Due to the resources involved in designing and implementing more robust, complex evaluations we would expect one or more **feasibility studies** to precede any full evaluation, with progression criteria clearly articulated. This should be accounted for in programme design, for example, using a pilot period to feasibility test the evaluation and develop appropriate measures, which can also be used to generate data on the pilot itself.

4. For reasons of methodological expertise and the need for objectivity, this means that **Tier 1 evaluations will usually be conducted by someone who is a qualified research or evaluation professional with relevant specific expertise in the evaluation of professional learning and organisational change**, and not members of the programme team or the GSCU evidence and impact team. Peer review and advice should be sought from the **Trials Advice Panel** via the Evaluation Taskforce.

5. Appropriate **statistical quasi-experimental methods are preferred**, where an appropriate data set exists or can be built, because they make fewer demands on colleagues' time, make effective use of existing resources, and contribute to building our overall DDaT capability.

6. In this tier it is particularly expected that attention will be given to identifying **robust and cumulative outcome measures**. That is, they will ideally be pre-existing, published, pre-validated measures, including having been validated for the specific context of the intended evaluation (as identified via appropriately robust external literature reviewing e.g. scoping review), and conceptually validated for the construct they are measuring.

7. The evaluation **must be led end-to-end by an internal analytical specialist** i.e. they are a badged member of an appropriate analytical profession.

8. Principles of Open Science should be followed, including, where relevant, pre-registering evaluations using an appropriate platform, and ensuring findings are made publicly available.

Design Principles for Tier 2 evaluations:

1. Well-conducted **Rapid Evidence Assessments (REAs)** of appropriate external academic and 'grey literature' evidence, would generally be most appropriate at this tier for synthesis of external evidence. Nonetheless, for innovative interventions or where the evidence base is not clear, a greater level of evidence synthesis may be required. Thought should be given to thematic connections between interventions (e.g. multiple interventions targeted at the same skill area), where reviews of evidence targeted at that strategic theme may be more efficient than undertaking multiple smaller reviews.

2. Tier 2 evaluations will generally entail the roll out of **standardised data collection tools and approaches**, tailored to specific groups or clusters of related Campus programmes/activities, which are **implemented by delivery teams as part of their day-to-day work**, with analysts and/or occupational psychologists leading the design, analysis and generation and interpretation of findings. Input of external evaluators or specialists may be required, but in a more targeted way, for example to support the development of methods and measures, or during a pilot phase where more intensive evidence may be sought to optimise the design long-term.

3. Within Tier 2, as far as possible, **existing administrative/MI data** should be used and where additional data is collected, **clear benefit of the intended data collection to the colleagues involved in collecting and supplying it** should be evident.

4. Tier 2 evaluations will be more explicitly **co-designed/delivered between Campus analysts/occupational psychologists and delivery teams**, to get the right balance between rigour and maximising insight, and the 'implementability' of the data collection tools and approaches. Tier 2 evaluations will generally **not require fully independent evaluation design and delivery, i.e. programme delivery teams (including external suppliers from whom an intervention has been procured) are likely to be involved in all or some aspects directly**. Independent scrutiny and peer review is particularly important in this situation, for example from the Evaluation Taskforce.

5. Tier 2 evaluations in particular **must have educative benefits for the colleagues involved**. That is, thought has been given to how non-analytical colleagues will learn and benefit from being involved in the evaluation process, and this has been embedded in the design. This may include, but would not be limited to, through processes of participatory co-design, through educative data collection tools that generate new thinking skills, through mechanisms of professional reflection, through participation in data analysis and interpretation as a co-researcher, or through processes of dissemination of and action on findings.

6. Our existing networks of academic expertise should be leveraged as far as possible to support research and evaluation in this tier.

7. It will generally not be proportionate in Tier 2 to conduct experimental designs, but quasi-experimental approaches may still be viable, dependent on the level of resources required to access data and create a usable data set and/or to identify appropriate outcome measurement tools for pre/post testing. **Resources devoted to Tier 2 evaluations are expected to be lower than those devoted to Tier 1**, so it would not be proportionate to carry out a resource-intensive evaluation at this level. Other methods that it will be appropriate to consider include other statistical methods for analysing numerical data, pre/post testing methods using appropriate outcome measures which have been internally well-designed but not pre-published/validated, and high quality qualitative approaches (e.g. observational methods, interviews, focus groups), including imaginative use of existing administrative data (e.g. document analysis).

Design Principles for Tier 3 evaluations:

1. Formal evaluation methods are unlikely to be needed within this tier, and are discouraged because it would not be a proportionate use of resources. Over-evaluating programmes also creates the risk of fatigue among learners and managers due to an excessive number of 'asks' for data, which will impair evaluation in Tiers 1 and 2.
2. Monitoring/MI data should still be undertaken within this tier, of a selected number of crucial data points - for example, spend, participation rates, learner profile, quality of experience - with ongoing processes being in place to check these are at expected levels and explore any discrepancies. **Standardised MI approaches should be used as far as possible**, that is, all interventions should collect the same data, as specified by the Campus Evidence and Impact team. Additional data can also be collected where relevant to do so.
3. If you feel there is a case for formal evaluation of a Tier 3 programme this can be discussed with the GSCU analytical team.