May/2026

HM Treasury

Transparency in Government Evaluation Research (TIGER)

March 2026
Evaluation Task Force

This guidance was drafted by Jack Blumenau (University College London), Sarah Nila (University College London), and Anouk Rigterink (Durham University) on behalf of the Evaluation Task Force. The authors gratefully acknowledge support from UK Research and Innovation for funding this work (grant number: UKRI1081).

Table of Contents

Foreword

The primary goal of policy evaluation is to sharpen our understanding of what works, what doesn't, and why. For evaluations to effectively inform government decision making, they must be grounded in evidence that is both robust and reliable.

In recent years, the global scientific community has made significant strides in strengthening research quality through the adoption of open science principles and practices. Transparency is not merely an academic aspiration; it is a necessity for generating high-quality evidence that improves how we govern. By being open about how we plan, analyse, and report evaluations, we safeguard the integrity of the evidence we generate, prevent waste, and raise standards.

This guidance offers a clear framework for applying open science principles to government evaluation. It challenges us to "show our working", from initial design through to final analysis. The practices set out here – such as detailed analysis plans, open data and code, and structured reporting standards – help ensure that we do not shift the goalposts when evaluating policy effectiveness, and that we are as transparent about our uncertainties as we are about our findings.

Crucially, this guidance is built on the principle of proportionality. It recognises the importance of balancing transparency and rigour with the practical realities of government timescales, operational constraints, and resource pressures.

Adopting these practices will accelerate the shift in culture toward greater transparency that is already underway. The benefits are significant: improving transparency will strengthen the quality of evidence, enable better scrutiny, and ultimately support better outcomes for the public. This guidance represents a vital step toward a more open, rigorous, and accountable evaluation ecosystem. We are delighted it has been made possible by a UKRI Metascience Research Grant.

Jen Gold,
Executive Director of Research, Economic and Social Research Council, UKRI

Stian Westlake,
Executive Chair, Economic and Social Research Council, UKRI

Glossary

False Negatives: False negatives are results suggesting that an intervention had no effect, when in reality it did.

False Positives: False positives are results suggesting that an intervention had an effect, when in reality it did not.

Hypothesising After the Results are Known: Presenting a hypothesis developed after analysing the data as if it were specified beforehand.

Open Access: Making research outputs freely available to the public without paywalls or access restrictions.

Open Code: Making the code used to generate research findings freely accessible for scrutiny and reuse.

Open Data: Making raw research data freely accessible so others can verify findings or reuse the data for further analysis.

Open Materials: Sharing research instruments (such as surveys, stimuli, code) used in a study to enable replication and reuse.

P-hacking: Manipulating data or analyses (for example, trying multiple statistical tests or stopping data collection early) to obtain statistically significant results.

Power Analysis: A statistical method used to determine the minimum sample size required to reliably detect a true effect of a given size, or to estimate the smallest effect that can be confidently detected with a specific sample size.

Preregistration: Publicly recording the research plan, hypotheses, analysis strategy, sample size calculations, and power analyses before collecting data to improve transparency and reduce bias.

Publication Bias: The tendency for studies with positive or statistically significant results to be published more often than those with null or negative findings.

Replication: The independent verification of research findings by collecting new data and repeating the original methods and procedures of the initial study to see if the same conclusions are reached.

Reproducibility: The ability of an independent researcher to recreate the original results of a study using the same original data, code and computational procedures provided by the first researchers.

Robustness: The extent to which the results of a study remain qualitatively similar when subjected to reasonable changes in the data analysis, such as using different analytical methods or alternative model specifications.

Selective Outcome Reporting: Reporting only a subset of the originally measured outcomes, typically those that are statistically significant, while omitting others.

Introduction

This guidance champions transparent research practices for government evaluators. It adapts key principles of the 'open science' movement, which aims to make research more transparent and accessible, for the government context.

Open science is a term for a set of practices that aims to make scientific research and its outcomes more accessible to external parties. Research is 'open' when the parts of the research process that are normally known only to individual researchers or teams are made more broadly available to others. Transparent practices in evaluation include documenting and preserving evaluation plans (such as evaluation protocols and statistical analysis plans), evaluation materials (such as data and code), and the clear reporting of evaluation findings.

Transparency serves to strengthen the integrity of, and trust in, evaluation findings. Transparency also maximises value for money and promotes efficiency by enabling the reuse of code, data and other materials, and promotes effective relationships with suppliers. Ultimately, transparency in evaluation ensures that policy decisions are underpinned by evidence which is credible, trustworthy and open to scrutiny, enhancing the confidence of political audiences and the public in government evaluations.

This guidance supports the consistent and proportionate application of transparency in government evaluation by providing examples of what evaluation transparency can look like in practice throughout the evaluation process – from planning, to analysis, to reporting.

This guidance provides nine "good practice" recommendations to help evaluators make their research more transparent.

Many practices to promote transparency apply to both quantitative and qualitative methods of evaluation. This guidance clarifies where and how each practice applies to quantitative and qualitative research.

Proportionality

Recognising that one size does not fit all, these standards should be applied proportionally to the value, risks, costs and effort involved in each individual evaluation, with the most stretching practices more likely to be applied in higher stakes evaluations or where the benefits of transparency are likely to be greater. In practice, this means that evaluations which:

- Are high cost.
- Evaluate high-profile or high-cost policy interventions.
- Have high learning potential or potential for re-use, or.
- Use novel or complex methodological approaches.

should be subject to greater transparency. Legal constraints and constraints relating to data sensitivity may also shape the choice of transparency practices.

Research informing this guidance

This guidance was informed by a review of existing resources on transparency best-practices from a wide variety of research-producing organisations, both in the UK and internationally. The review included 123 documents and synthesised recommendations from these sources. We also conducted interviews with 11 evaluation stakeholders from six government departments and arms-length bodies to understand existing practice.

Relation to other guidance

This annex complements existing frameworks that support transparency in UK government research, by providing practical, proportionate recommendations across all major steps of evaluation.

1. The Government Social Research Publication Protocol sets out clear standards for publication transparency, outlining responsibilities for releasing research outputs, maintaining independence and ensuring timely dissemination.
2. The Code of Practice for Statistics articulates high-level principles of transparency.
3. The Analytical Quality Assurance (AQuA) Book provides guidance for producing quality analysis in government and also emphasises the importance of transparency.
4. The Concordat to Support Research Integrity provides a framework for good research conduct and its governance.

This annex focuses on how to operationalise the principles of open science in the context of government evaluations, while sharing the goal of promoting research transparency.

Benefits of adopting transparent evaluation practices

The adoption of transparent research practices can benefit government evaluation in several ways.

1. **Strengthening the integrity of evaluation research**

   Open science research practices were developed in response to concerns about the reproducibility, quality and scientific integrity of research across many academic disciplines. These concerns have arisen in response to research that has documented difficulties replicating the results of prior studies in different contexts; reproducing results even when using the same data in the same context; and weaknesses in the robustness of results, where small changes in analytical choices can lead results to change substantially.[1] These failures have been attributed to a number of questionable research practices, including insufficient statistical power, flexible rules for stopping data collection, selectively reporting outcomes or treatment conditions, hypothesising after the results are known and manipulating data or analyses to obtain statistically significant results. (such as p-hacking).[2]

   There is growing evidence that the adoption of open science research practices can help to mitigate these issues and research transparency is seen as critical to enhancing the credibility and societal value of research.[3] The adoption of open science practices should therefore both raise the reliability of evaluation findings, and strengthen the robustness of the evidence base used in government decision-making.

2. **Value for money, efficiency and reuse**

   Transparent practices can improve the value for money of evaluations. Such practices facilitate the development of further useful analyses which may have been unforeseen at the time of the original evaluation. They may also facilitate the re-use of code or data. This increases value for money and promotes efficiency by avoiding duplication of efforts.

1 See, for instance, (Open Science Collaboration, 2015; Chang and Li, 2022; Obels et al., 2020; Breznau et al., 2022).
2 See (John et al., 2012; Stefan and Schönbrodt, 2023; Ioannidis, 2005; Chan et al., 2004; Kerr, 1998; Simmons et al., 2011).
3 See (Nosek et al., 2012; Dudda et al., 2025; Brodeur et al., 2024; Hardwicke et al., 2018; Blanco et al., 2020; Cobo et al., 2011) for examples.

3. **Promoting effective working relationships with suppliers**

   Adopting transparent practices can also help government work effectively with suppliers of externally commissioned evaluations. For example, a detailed analysis plan, authored by either the commissioning government body or an external supplier, creates clarity on what work commissioned evaluators are expected to do. Requiring the supplier to share data and code with the government body enables scrutiny of the supplier's work, and ensures data and code can be reused and is not lost.

4. **Building public trust in evaluation**

   Adopting transparent research practices could also improve public trust in government evaluations. Encouraging analysts to be more transparent enables independent scrutiny and replication of evaluation findings. This practice aligns with the Code of Practice for Statistics, which suggests that transparency has an important role to play in supporting the pillars of trustworthiness, quality and value in public statistical research.

5. **Equity and accessibility**

   Transparency also ensures that the outputs of publicly funded evaluations are accessible to all, regardless of institutional affiliation or location. This widens the audience for government research, maximises the public value of evaluation, and promotes fairness in how evidence is shared and used.

Together, adopting an open approach to evaluation in government is likely to have a number of beneficial effects. By embedding transparency throughout the research life cycle, departments and analysts can demonstrate methodological rigour, invite constructive challenge, and, ultimately, deliver more trustworthy evidence for policymakers and the public.

Principles of research transparency in evaluation

This guidance organises practices according to principles that align with different parts of the evaluation process: planning, analysis and reporting. Underpinning transparency practices at each stage is a cross-cutting proportionality principle.

Principle 1: Proportionality

Evaluations should apply transparent research practices in a proportionate way, weighing the value and benefits of transparency against risks, costs and effort.

Principle 2: Planning transparency

Evaluations should be planned in a transparent manner. Planning transparency involves documenting the intent, design and methodology of an evaluation before it begins. For quantitative evaluations, transparent planning also involves conducting and reporting power analyses. Evaluation plans should be kept for future reference within departments and, where proportionate to do so, these details should be shared in a way that allows them to be scrutinised by others. Planning transparency creates a record of intended evaluation approaches, thereby strengthening an evaluation's integrity.

Principle 3: Analysis transparency

The analytical processes used to produce evaluation findings should be transparent. Analysis transparency involves preserving any materials, data and code used to generate an evaluation's findings and, for quantitative evaluations, ensuring that final results are computationally reproducible. These materials should be securely stored within departments to support future use and, where proportionate to do so, should be shared in a way that allows them to be scrutinised by others. Analysis transparency enables the independent verification of evaluation results, strengthening confidence in their robustness and integrity and promoting value for money through reuse.

Principle 4: Reporting transparency

Evaluation reports should communicate findings in a complete and accessible manner. Reporting transparency involves clearly and fully describing all key methodological choices, communicating the uncertainty around findings and demonstrating the robustness of results to alternative analytical approaches. Reporting transparency builds trust and facilitates better-informed policy decisions by giving users a comprehensive understanding of the evidence base and its limitations.

The remainder of this guidance includes further details on how these principles should be applied in government evaluations in practice.

Summary of recommendations

Table 1: Summary of recommendations.

Principle	Recommendations
Planning	**1. Register evaluations.** All planned, live and completed evaluations should be registered on the Evaluation Registry. **2. Timestamp and archive plans**. Evaluation plans, protocols *and statistical analysis plans* should be finalised, time-stamped and archived internally, before data collection or analysis begins. Where proportionate to do so, these documents should also be published (for example, on a registry or as an annex) or deposited with the Evaluation Task Force in advance of analysis. **3. Calculate power.** All statistical analysis plans should include power calculations.
Analysis	**4. Share materials.** Materials (such as survey instruments, questionnaires, codebooks, etc.) used in an evaluation should be attached as an annex to the final evaluation report. **5. Use and preserve code.** Data analysis should be performed using code written in an open-source language (such as R, Python and so on) where possible. Code should be archived internally so that it is preserved and accessible to others in the department. Where proportionate to do so, code may also be stored in a repository that allows for external access once the evaluation is complete. **6. Archive data.** Evaluation data should be archived internally within the department. Data availability statements, which set out who owns evaluation data and arrangements for access, should be included in evaluation reports. Where proportionate to do so, evaluation data may also be stored in a repository that allows for external access once the evaluation is complete.

Principle	Recommendations
Reporting	**7. Follow reporting guidelines.** Evaluation reports should follow established reporting guidelines to ensure analysis choices are explained completely and transparently. **8. Communicate uncertainty.** Evaluation reports should present measures of statistical uncertainty (such as confidence intervals, standard errors) alongside point estimates and explain what these imply in accessible language. **9. Conduct and report sensitivity analyses.** Analysts should conduct appropriate robustness checks, testing core findings against alternative analysis choices with results reported in the evaluation report.

Note: Recommendations in *italics* apply to evaluations using quantitative methods only. Recommendations not in italics apply to evaluations using either quantitative or qualitative methodological approaches.

Proportionality in transparent evaluation practices

While the benefits of transparent research practices are likely to apply to a wide set of government evaluations, not all evaluations require the same level of transparency.

Greater attention to transparency should be considered for evaluations with the following characteristics.

1. Evaluations that are high cost.
2. Evaluations of high-profile or high-cost policy interventions.
3. Evaluations that have high learning potential or potential for re-use of data, code or materials.
4. Evaluations that employ novel or complex methodological approaches.

For example, an evaluation team conducting **a small-scale evaluation of a low-cost intervention, where the data and code produced by the evaluation are unlikely to be reused, might follow lighter-touch transparency practices**. In such a case, the team might produce a short protocol and statistical analysis plan to be stored internally; ensure that data and code produced by the evaluation is saved in an appropriate place within the department; and clearly explain any analytical choices and measures of uncertainty in the final evaluation report.

By contrast, an evaluation team conducting a **high-cost evaluation of a high-profile intervention might follow more stretching transparency recommendations**. In such a case, the team might ensure that the protocol and statistical analysis plan are registered in a secure research repository in advance of data collection; that code, data and any other materials are made available to external researchers; and that the evaluation contains a full suite of robustness checks which are clearly communicated in the evaluation report.

Evaluators should thus balance the need for transparent research with practical constraints, applying the appropriate level of openness based on the project's characteristics.

Planning transparency

Planning transparency ensures that the intentions, design and analytical approach of an evaluation are documented in advance of any analysis being conducted. Practices to ensure planning transparency include study registration, evaluation plans and protocols, statistical analysis plans and power calculations.

Planning transparency serves a number of **purposes**.

1. It can strengthen accountability by creating a record of the evaluations that are planned or in progress, thereby encouraging the publication of those evaluations once they are complete.
2. It can facilitate planning by encouraging teams to specify in detail how the evaluation will be delivered, reducing the risk of later implementation problems.
3. By pre-specifying details in an evaluation protocol and statistical analysis plan, analysts can increase the analytical integrity of research by fixing key analysis decisions in advance of data collection and analysis. This is vital for reducing the scope for adjustments after results are produced that could bias or misrepresent evaluation findings.
4. Pre-specifying plans can insulate analysts from external pressures to adapt methodological strategies in response to unflattering evaluation findings.

However, it is possible that plans have to be adapted. Evaluations are iterative or may face changing policy or other needs during their delivery and it is often important for plans to evolve in response to unforeseen circumstances (such as issues with data collection, or changes to the policy or intervention being evaluated). In these instances, analysts should not rigidly stick to planned analyses that are no longer workable. If the plan must be adapted to achieve the best possible analysis, evaluations should record the change made and the rationale behind it – either by updating the relevant documents or by clearly explaining the decision in the final evaluation report.

Study registration

Study registration is the act of creating a public record that an evaluation exists or is about to begin. Study registrations can include minimal levels of detail that capture only the essential aspects of an evaluation, such as the intervention title, the department responsible for running the evaluation, high-level design features (such as whether the evaluation is an impact, process or value for money evaluation), and expected publication dates. The purpose is to make clear the intent to evaluate a particular intervention rather than to describe the specifics of the evaluation in detail.

The main purpose of a study registration is that it creates a simple commitment device to ensure that evaluations that are planned and implemented are not subsequently prevented from being released.

Box 1: The Evaluation Registry

The Evaluation Registry is a website which acts as a central repository for all government evaluations. An important requirement of the registry is that all planned evaluations – not just ongoing and completed evaluations – must be registered after the evaluation has been approved by the relevant internal governance processes. This ensures that there is a record of approved evaluations, meaning that the departments responsible for producing those evaluations can be held to account if they are not subsequently published.

Registering a planned evaluation on the Evaluation Registry requires minimal detail, with researchers only required to provide the title of the evaluation, the type of evaluation, a high-level description of the intervention being evaluated, and an indication of expected publication dates. At the time of writing, the Evaluation Registry contains more than 1,400 completed evaluations, and more than 400 planned and ongoing evaluations.

Recommendation

1. Register evaluations. All planned, live and completed evaluations should be registered on the Evaluation Registry.

How to apply this recommendation in government evaluation?

All central government departments are required to register all planned, live and completed evaluations on the Evaluation Task Force's Evaluation Registry (see Box 1). Departments are required to register planned evaluations 'as early as is practicable' under the terms of the Government Social Research publication protocol, no later than 12 weeks after sign-off by the relevant internal departmental governance processes, and before the first round of data collection for the evaluation takes place.

Evaluation plans and protocols

Evaluation plans document the questions an evaluation seeks to answer, the data and methodological approach being used, and the plan for delivering the evaluation. The level of detail specified in an evaluation plan will vary depending on the nature of the evaluation. For some evaluations, the plan will include only relatively high-level details on study goals and objectives, key research questions, broad methodology, timelines, governance arrangements and ethical approval. For other evaluations – particularly quantitative impact evaluations using randomised controlled trials or quasi-experimental methods – evaluators might provide much more detail by documenting their plans in a formal evaluation protocol.

Evaluation protocols provide a detailed blueprint of the design of an evaluation and explain how the work will be delivered. The typical contents of a protocol include:

- A Theory of Change for the intervention.
- Evaluation questions.
- Details on the sampling and recruitment strategy for any participants.
- Data sources.
- Data-collection timelines.
- Issues related to governance, risks and ethics.
- Details on the procedures used to allocate units into different treatment conditions (in the case of randomised trials).
- Sample size calculations and power analyses (for quantitative evaluations, discussed below).

Documenting design choices in an evaluation protocol is useful for guiding the future delivery of evaluations because it creates a single resource which all stakeholders can use as a reference throughout the delivery process.

Both evaluation protocols and the statistical analysis plans (see below) should clearly distinguish between **confirmatory and exploratory research questions**. A confirmatory research question is one that has a clearly articulated hypothesis and outcome(s), established before any data are examined, making transparent how the success of an intervention will be judged. By contrast, an exploratory research question is one that aims to discover unanticipated patterns or generate new hypotheses after or while viewing the data. This signals that any patterns revealed are provisional and will often require further research before they are used to inform decision-making.

Helpful tools

Evaluators drafting evaluation plans and protocols can draw on existing checklists and guidance to ensure that their documents are comprehensive.

1. The SPIRIT 2025 statement provides a widely used framework for the minimum wcontent of trial protocols, with an accompanying SPIRIT checklist that can be adapted for government evaluations.
2. The Education Endowment Foundation also provides protocol templates for both randomised controlled trials and quasi-experimental evaluations.
3. The Transforming Access and Student Outcomes (TASO) What Works Centre also provides a protocol template for qualitative evaluations.
4. The Rainbow Framework is useful for all single and multi-method evaluations.

Box 2: Evaluation plans: Office for Value for Money

The Office for Value for Money (OVfM) evaluation aimed to assess the effectiveness of a new multidisciplinary team within HM Treasury aimed at reducing waste and improving efficiency. The evaluators produced an evaluation plan for a proportionate, light-touch evaluation of the OVfM. The evaluation plan included:

1. A clear set of evaluation aims covering process, impact and value for money questions.
2. A high-level description of evaluation methods, including the use of document reviews, anonymous stakeholder surveys and in-depth interviews.
3. A set of evaluation success measures, aligned to the process, impact and value for money evaluation questions.
4. A targeted sampling strategy identifying specific stakeholders for feedback.

By documenting these choices in a published plan, the evaluators created a clear 'roadmap' for the evaluation. This ensured that key elements of the plan – such as the criteria for success – were fixed in advance, preventing the goalposts from shifting at a later date. However, in contrast to a more detailed evaluation protocol (see Box 3 below), the plan did not commit the evaluators to a precise methodological approach, meaning that they retained flexibility in the ways in which the broader plan was implemented.

Box 3: Evaluation protocols: Health-Led Trials

The Health-Led Trials evaluation investigated the effects of an Individual Placement and Support (IPS) service which aimed to provide employment support to individuals with mild to moderate mental and/or physical health conditions. The evaluators wrote an evaluation protocol which included:

1. A statement of the research questions which were the focus of the evaluation and a description of the expected effects of the trial.
2. A description of the intervention.
3. Definitions of each of the primary and secondary outcomes used to evaluate the success of the intervention.
4. Information on the composition of the sample selected for inclusion in the trial.
5. An outline of publication plans for the evaluation.

Each of these components helped to define the scope of the evaluation and committed the evaluators to a specific evaluation approach. The protocol was published in advance of the intervention being rolled out, meaning that the results of the evaluation could not influence the evaluation design. The evaluation was also adapted at several points throughout the evaluation process, with changes to the design clearly described in amendments to the protocol.

Statistical analysis plans

Statistical analysis plans are technical companions to evaluation protocols for quantitative evaluations that specify exactly how data will be cleaned, processed and analysed, before analysts access the data. The vast range of analytical choices that a researcher needs to make in conducting any quantitative analysis means that researchers are able to produce a wide range of impact estimates from the same underlying data. Pre-registration (see below) of a detailed statistical analysis plan encourages researchers to choose the analysis strategies that they think are most likely to lead to accurate estimates and prevents them from selecting analytical strategies on the basis of the results they produce.

While helpful to restrict after-the-event discretion, rigid plans can be counterproductive when data properties are unpredictable. For instance, a researcher might want to implement one type of analysis if attrition from a randomised-controlled trial is high and another type of analysis if attrition is low. In such cases, the statistical analysis plan can include 'what if' statements, detailing how the different analyses will be implemented and which decision-making rules will be used to decide on the appropriate strategy. By pre-specifying these decision rules, researchers can respond to the data's nuances while remaining insulated from the temptation to cherry-pick results based on the estimates they produce.

A statistical analysis plan should contain enough detail so a third party could produce the analysis without assistance. This might include information relating to:

- **Data transformation and preparation.**
  - Details of any planned data transformations between the original, raw data (for example, administrative, survey) and the final analytical dataset.
  - Specific steps for correcting anticipated data errors, imputing missing data, recoding outliers or aggregating variables into indices.
- **Analytical models and tests.**
  - The models or statistical tests (such as linear regression, logistic regression, fixed-effects, and so on) to be used to assess impacts on pre-specified outcome variables.
  - Specifications of which variables will be included in the analysis (both independent and dependent variables) and how they will be incorporated (such as linear or non-linear predictors, including interaction terms, and so on).
- **Statistical uncertainty.**
  - The planned approach for calculating and reporting statistical uncertainty, such as the use of clustered standard errors.
  - The method for addressing multiple hypothesis testing if the evaluation involves analysing multiple outcomes.

- **Assumptions, limitations and exclusions.**
  - Tests designed to identify potential limitations (such as balance tests, tests for selective attrition, and so on).
  - Pre-specified reasons and criteria for excluding individual observations or control variables (for example, due to excessive missing data).
  - Criteria that may lead analysts not to implement a pre-specified analysis (for example, failure of an assumption test or evidence of excessive missing data).

Helpful tools

1. Several checklists on what statistical analysis choices should be documented in a statistical analysis plan exist: see, for example, McKenzie (2012) and Wicherts et al (2016).
2. The Education Endowment Foundation provides a template for a statistical analysis plan.
3. Examples can be found on the Open Science Framework Registry.

Preserving and sharing planning documents

All planning documents – including evaluation plans, protocols and statistical analysis plans – should be completed before the evaluation begins, timestamped and preserved within the department. These documents should be made available to relevant stakeholders throughout the evaluation life cycle.

In addition to being stored internally, evaluation plans, protocols and statistical analysis plans may also be **pre-registered**. Pre-registration creates a time-stamped, immutable, publicly available record of the study design, which acts as an audit trail to protect the integrity of the analysis. The advantage of public pre-registration is that it allows stakeholders to verify that the study was conducted as originally intended, thereby minimising the risk of bias and significantly increasing confidence in the findings of a study evaluation.

While there are examples of published pre-registrations of government evaluations (see Box 3 for an example), public pre-registration may not be **proportionate** for government research. Specific constraints include:

1. Operational flexibility: Interventions and delivery timetables often evolve, necessitating updates to evaluation plans that make static pre-registration impractical.
2. Resource efficiency: The administrative burden of external publication processes may not represent good value for money for smaller or lower-stakes evaluations.
3. Sensitivity: Commercial, security, or data protection sensitivities may prevent the early release of detailed protocols.

Evaluators should therefore balance these constraints against the benefits of public pre-registration, specifically strengthening the integrity of, and public trust in, the evaluation.

Helpful tools

Public repositories for evaluation plans, protocols and statistical analysis plans include:

1. Open Science Framework Registry (OSF).
2. Social Science Registry.

Recommendation

2. Timestamp and archive plans. Evaluation plans, protocols and statistical analysis plans should be finalised, time-stamped and archived internally before data collection or analysis begins. Where proportionate to do so, these documents should also be published (such as on a registry or as an annex) or deposited with the Evaluation Task Force in advance of analysis.

How to apply this recommendation in government evaluation?

All evaluations produced in government should be accompanied by appropriate planning documents. The level of detail included in an evaluation plan should be proportionate to the scale and importance of the evaluation and should reflect the methodological approach used. For instance, for quantitative impact evaluations using either randomised controlled trials or quasi-experimental designs, it is good practice to write a detailed protocol using an established checklist. For evaluation activities that are primarily qualitative, or where specifying a full protocol would be disproportionate, it is good practice to write a higher-level evaluation plan (see Box 2 above for an example).

Plans and protocols should be developed early in the evaluation process and signed off by relevant governance bodies before data collection begins. Plans and protocols should be archived internally to ensure easy access for relevant staff. This allows evaluators and other stakeholders to refer to them, serving as a commitment device to uphold analysis choices, and helping to ensure continuity if project teams change. Protocols should be treated as live documents that can be updated if and when circumstances change, or if new information comes to light that has a bearing on the design of the evaluation. When this happens, deviations from the original protocol should be documented, allowing stakeholders to distinguish between planned and unplanned choices.

Quantitative evaluations should have a statistical analysis plan that is written before data collection starts (if new data is collected) or before data analysis starts (if no new data is collected). The level of detail included in a statistical analysis plan should be proportionate. For example, a statistical analysis plan for a low-cost, time-sensitive evaluation with one outcome that uses high quality administrative data and transforms this data in ways that have a strong precedent within a department, may count only a few pages. By contrast, the statistical analysis plan for a high-cost evaluation collecting new data, anticipating substantial issues with data quality, and analysing many outcomes will typically be longer. When reporting, any deviations from the statistical analysis plan should be described, and results from an analysis adhering as closely as possible to the statistical analysis plan should be presented, even if these are not the headline results.

Depending on what is proportionate, evaluators might take different approaches to time-stamping and archiving evaluation plans, protocols and statistical analysis plans, including:

- Archiving a time-stamped, immutable version of planning documents internally within the department.
- Sharing a time-stamped copy of documents with the Evaluation Task Force by emailing relevant documents to etfpreregistration@cabinetoffice.gov.uk. This approach is helpful where an extra layer of independent verification would strengthen the study's credibility, but where publication would be inappropriate or disproportionate.
- Publishing relevant documents in the public domain (such as on GOV.UK or a public registry).

Power analyses

A statistical analysis plan also often includes **power analyses**. Statistical power is the probability that an evaluation, using a given sample size, will successfully detect the effect of an intervention, assuming that effect is at least a specified minimum size. A power analysis is used to estimate one of two things: the minimum sample size needed to confidently detect a true effect of a specific size; or, the minimum effect size, an evaluation can confidently detect, given an assumed sample size.

Underpowered evaluations represent poor value for money, as they risk generating **false negatives**, which are results suggesting that a policy had no effect when in reality it did. This can lead to beneficial policy interventions being prematurely discontinued due to a lack of evidence, wasting the initial investment and missing positive policy outcomes. While power analyses typically focus on the risk of false negatives, an underpowered evaluation may also risk **false positive** results – claiming an effect exists when it does not – which can lead to the adoption of ineffective or even harmful policies. Underpowered studies are extremely common, even in published academic research. Power calculations therefore provide very important information, as they give analysts an estimate of whether an evaluation is underpowered prior to implementing it.

Conducting power analyses requires analysts to make several critical assumptions about key parameters which are inputs to the analysis. These inputs typically include the expected effect size of the intervention, the anticipated variability in the outcomes, the desired level of statistical significance, and the desired level of power. These choices must be clearly justified and documented in the statistical analysis plan.

Helpful tools

There are a range of tools available to support power calculations.

1. For simple designs comparing treatment and control group averages, user-friendly online calculators exist (such as J-PAL Power Calculator; EGAP Power Calculator).
2. For more complex designs, free and open-source software such as G*Power can handle multiple treatment arms, pre and post-intervention data, and a variety of statistical models.

3. Where designs are highly complex, power can also be assessed through simulation-based methods in software such as R or Stata (see the DeclareDesign package for R or the J-PAL sample size and power resources). Conducting power analyses using programmatic languages like R has the additional benefit of increasing the transparency by making it clear how the analysis was conducted (see Open Code section below).
4. Useful guidance on power calculations for more complex evaluation designs is provided by McConnell and Vera-Hernandez (2025).

Recommendation

3. Calculate power. All statistical analysis plans should include power calculations.

How to apply this recommendation in government evaluation?

All statistical analysis plans should include power calculations for the primary outcome(s) of interest. These should be proportionate to the scale and cost of the evaluation. At a minimum, this means using simple, established calculators or software with reasonable assumptions about effect size and outcome variability. For higher-cost or higher-stakes quantitative evaluations, power calculations should be based on careful consideration of likely effect sizes and outcome variability, drawing on previous studies, pilot data or administrative records where available. Where appropriate, calculations should be repeated under different assumptions to illustrate the trade-offs between sample size, cost and the risk of false negative and false positive results. Best practice also involves using simulations or more advanced methods for complex evaluation designs where simple calculators are insufficient.

Analysis transparency

Analysis transparency involves storing the materials (such as survey questionnaires, intervention protocols), data and code used to produce an evaluation so they can be accessed by others. Analytical transparency creates a complete record of how the analysis in an evaluation was conducted, enabling third parties to independently recreate the outputs of an evaluation.

In academia, materials, data and code are often made available by storing them in a **repository**. Some repositories store files so they can be accessed by anyone, while some have more limited access, where materials can be accessed only by trained and accredited researchers. For government research, in many cases it may be more appropriate for materials, data and code to be stored internally with availability restricted to only those in a given department or evaluation team. The decision on the level of access will depend on how sensitive materials, data and code are.

Analytical transparency serves a number of **purposes**. It can:

- Strengthen the integrity of, and trust in, evaluations, by ensuring that they are computationally reproducible.
- Enables scrutiny of the evaluation by enabling external parties to see in detail how the evaluation was conducted.
- Help increase the value for money of an evaluation by making it easier to do further analyses, and to exploit synergies across evaluations.

Open materials

Open materials refers to the practice of making materials used to collect data or that constitute a treatment for an evaluation, available to others. For quantitative evaluations, this might include survey instruments or questionnaires used to collect primary data; codebooks detailing the structure and content of data files; and any resources (such as videos, images, pamphlets) that constitute the content of an intervention. For qualitative evaluations, this might include interview instruments, such as topic guides used for individual interviews or focus groups; codebooks which detail how textual data (transcripts, documents) was coded; images, videos or documents shown to participants to elicit discussion; and any researcher training manuals or protocols. Most commonly, such materials may be made available to others as an annex to an evaluation protocol, or in an evaluation report.

Recommendation

4. Share materials. Materials (such as survey instruments, questionnaires, codebooks and so on) used in an evaluation should be attached as an annex to the final evaluation report.

How to apply this recommendation in government evaluation?

Unless sensitive or protected by intellectual property rights, materials used to produce an evaluation should be published when an evaluation report is published. If not provided in the evaluation report, the report should make clear where these materials may be found.

When materials are sensitive or protected by intellectual property rights, the evaluation protocol or report should justify this in a materials availability statement. This statement should also specify who (if anyone) may gain access to the materials and under which conditions. Access to materials should be restricted only so far as the sensitivity of the materials and/or intellectual property rights require.

Open code

Open code refers to the practice of releasing the computer scripts or other executable files that turn raw data into the tables, figures and statistical outputs presented in an evaluation report. These scripts include every step of the analytical pipeline, including how the data are cleaned, how variables are constructed, which statistical models are fitted, and how results are visualised.

Creating code rather than using manual steps to analyse data, such as copy-paste, point-click or drag-drop operations, has substantial advantages.

1. Using code is essential for ensuring that evaluation analyses are computationally reproducible. An evaluation is reproducible if others can independently obtain the same results using the same data. Lack of reproducibility is surprisingly common even in published academic studies[4] and can severely undermine the perceived integrity of an evaluation.
2. Using code provides analytical flexibility: should analysts want to change something early in the analytical workflow, they can do so without having to manually replicate later steps.
3. Using code enables the analyst or others to easily replicate (parts of) the analytical workflow to benefit a different evaluation, increasing value for money.

The value and utility of code are magnified when the following practices are adopted:

- **Accessibility:** Making the code public, or at least easily shareable, significantly enhances transparency and the possibility of external scrutiny.
- **Clarity:** Ensuring that code is well-commented allows a third party to quickly understand the purpose and functionality of specific lines and sections.
- **Open-source software:** Building code using open-source software (such as R and Python) ensures the analysis is accessible to anyone.
- **Version control:** Using a system to manage different versions of the code ensures an audit trail of analysis choices.

Helpful tools

Several repositories for code exist, including:

1. UK Data Service ReShare Repository.
2. Open Science Framework repository.
3. GitHub.

4 (Chang and Li, 2022; Obels et al., 2020).

Recommendation

5. Use and preserve code. Data analysis should be performed using code written in an open-source language (for example, R, Python and so on) where possible. Code should be archived internally so that it is preserved and accessible to others in the department. Where proportionate to do so, code may also be stored in a repository that allows for external access once the evaluation is complete.

How to apply this recommendation in government evaluation?

Data analysis for quantitative evaluations should be done by means of code, which should minimise manual steps and should be well-commented.[5] Where manual steps are necessary, these should be clearly documented. Where possible, code should be written in open-source software. If there are multiple versions of code, these should be stored individually, with clear indications of version in file names, and a version description in the code itself.

Code, including any code created by a supplier, should be archived internally to ensure easy access for relevant departmental staff after data analysis has been completed. Where it is proportionate to do so, code for quantitative evaluations should also be published in a public repository. Code should not be published in cases where materials are sensitive or where they are protected by intellectual property rights.

5 For further guidance, see (Government Analysis Function, 2026) Reproducible Analytical Pipelines (RAP) strategy.

Box 4: Open code: ONS Health Research Group evaluations

The Office for National Statistics' Health Research Group evaluated the effects of bariatric surgery on patients' labour market outcomes. The researchers used multiple sources of administrative data and fixed-effects regression modelling to estimate average changes in employee pay and employment status after bariatric surgery. Although the evaluation report provides only high-level methodological detail, the researchers published full analysis code on a public Github page.

Written in the open-source programming language, R, the code documents each stage of the analysis process, including data cleaning and merging; model specifications; visualisations of results; and sensitivity analyses. The open release of this code provides an effective way for others to scrutinise the approach taken by the researchers. It also provides a valuable template for future work by other analysts. Similar code for other quantitative projects published by the team is available on the Health Modelling Hub Github page.

Open data

Open data refers to the practice of sharing the underlying datasets used to produce an evaluation. Open data also requires sharing appropriate documentation and metadata which ensures that datasets can be independently used and understood by others.

The UK government is a signatory to the Open Data Charter, which includes **an expectation that all government data will be published openly by default**. However, if data includes sensitive information, it is not always reasonable or proportionate to make the data used for government evaluations fully open. Depending on the sensitivity of the underlying data, evaluators may therefore wish to adopt different approaches. The appropriate level of openness should be determined prior to data collection with individual participants, so that consent statements can reflect that data will be made open to some degree.

1. **Fully open data.** Making data fully open involves posting the data so that it can be accessed online without restrictions. Fully open data is appropriate in cases when the data is non-sensitive and non-disclosive and where sharing data openly is consistent with relevant legal frameworks, such as the UK's data protection legislation. Existing repositories that can store data include the UK Data Service ReShare Repository, the UK Data Archive Qualibank, the Open Science Foundation (OSF) registry and the Harvard Dataverse.
2. **Partially open data.** Partially open approaches involve storing sensitive data in secure research environments where access is restricted to accredited researchers under defined conditions. Partially open data is appropriate when the data contains personal, commercial, or otherwise disclosive information that cannot be made public. Data can be released in a partially open way via platforms such as the UK Data Service ReShare Repository, the UK Data Archive Qualibank, ONS Secure Research Service (SRS) or Integrated Data Service (IDS), or other departmental repositories (such as the HMRC Datalab).
3. **Restricted data access.** In some cases, the sensitivity of the data or legal and ethical constraints mean it is appropriate to restrict access to analysts within a particular department or research team. However, it remains important that departments preserve an exact copy of the data used in evaluations within an internal folder or repository, so it can be used to revisit or replicate evaluation results. It is important that any data is stored in a persistent fashion so that it is not lost when individuals or teams involved in the evaluation change.

Decisions on data openness should be applied in line with existing legal requirements and departmental procedures and practices. In particular, evaluators must comply with the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018, and any departmental records management and retention policies.

Recommendation

6. Archive data. Evaluation data should be archived internally within the department. Data availability statements, which set out who owns evaluation data and arrangements for access, should be included in evaluation reports. Where proportionate to do so, evaluation data may also be stored in a repository that allows for external access once the evaluation is complete.

How to apply this recommendation in government evaluation?

Regardless of the level of openness, or type of data, all departments should ensure that evaluation data is preserved in ways that ensure its long-term usability. Where open data access is not proportionate or feasible, all datasets used to produce evaluations should be archived internally to preserve their use for analysts in the future. Data should be stored in machine-readable formats wherever possible, and data files should be appropriately versioned so that updates or corrections do not overwrite the original evaluation dataset.

Many evaluations rely on data collected by external contractors, suppliers or research consultancies. In these cases, departments should set clear expectations from the outset about data ownership, transfer requirements and transfer mechanisms. Where appropriate, tenders should specify that raw data must be returned to the commissioning department at the conclusion of the project. All evaluations should include a short data availability statement that sets out who owns the data used in the evaluation and what arrangements exist for others to access it.

Where it is proportionate to do so, data should be made fully or partially open to external users. Datasets should be accompanied by clear documentation and metadata to provide users with information that allows them to independently understand the data, enable discovery and allow and encourage research re-use. Documentation should include where and how the data was collected, and provide clear descriptions of each of the variables and value definitions included in the data. In addition, a data availability statement should be included in the final evaluation report to make clear whether and where the data used in the evaluation can be accessed.

Box 5: Open data: Evaluation of the Evaluation Academy and the Education Endowment Foundation Archive

Fully open data: In a randomised controlled trial of the Evaluation Academy – a training programme aimed at improving the evaluation skills of civil servants – the Evaluation Task Force produced a fully open replication dataset. This dataset, hosted on the UK Data Service ReShare Repository, included information and code required to replicate results presented in the evaluation report. The data was prepared to be non-disclosive by masking information that could identify participants. Participants were informed about the plan to publish data. Making this data fully open provides maximum transparency and facilitates reanalysis and replication.

Partially open data: The Education Endowment Foundation (EEF) – a What Works Centre – is an independent charity that works to provide evidence on the effectiveness of interventions in schools, colleges and early years settings. The EEF archives all the data from impact evaluations that it commissions, with data from almost 140 EEF-funded evaluations deposited in the archive since 2011. The EEF Archive is hosted in the Secure Research Service (SRS) of the Office of National Statistics (ONS). This data is partially open, as access to the archive requires approval by the EEF and other relevant stakeholders such as the ONS and the Department for Education. Researchers also have to be accredited via the ONS 'accredited researcher' scheme. The archive enables crucial evidence-building on education interventions while securely safeguarding sensitive data from schools and participants.

Reporting transparency

Reporting transparency involves communicating the results of evaluations in a clear and structured manner, ensuring that findings are presented with appropriate levels of confidence. It also requires evaluators to ensure that the findings they report are robust to reasonable alternative methodological choices. Reporting transparency increases the credibility and trustworthiness of evaluation evidence.

This section covers three interrelated aspects of **reporting transparency**:

1. The use of reporting guidelines ensures that evaluation reports provide a clear and faithful account of the analysis, including detailed descriptions of methodological choices, which provides readers with the information necessary to critically assess an evaluation's strengths and weaknesses.
2. Clearly communicating uncertainty in evaluation findings ensures that quantitative evaluations are presented with appropriate caution, avoiding the impression of unwarranted certainty.
3. Transparent reporting of assessments of robustness demonstrates whether quantitative evaluation findings are sensitive to alternative analytical choices.

Reporting guidelines

Reporting guidelines outline a structured approach to writing evaluation reports so they provide a clear and complete account of how an evaluation was conducted. These guidelines provide detailed recommendations on how to describe the intervention of interest, the methodological approach employed, the construction of treatment and control groups, the outcomes used in the analysis, data collection instruments, missing data, statistical methods and research ethics. These checklists help evaluators to report their evaluations completely and transparently, ensuring readers have the information they need to critically appraise, interpret and use the research.

Helpful tools

There are pre-existing reporting guidelines for many different types of study design, including RCTs, QEDs and studies using qualitative methods.[6]

1. The Consolidated Standard of Reporting Trials (CONSORT) cover randomised controlled trials.
2. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines provide information relevant to observational and quasi-experimental studies.
3. Standards for Reporting Qualitative Research (SRQR) cover qualitative evaluation.

- Reporting guidelines for specific evaluation methods such as propensity score matching[7] and economic evaluations (such as the Consolidated Health Economic Evaluation Reporting Standards (CHEERS)) also exist.

Departments may wish to develop their own reporting guidelines which standardise how evaluations are reported on. For example, the Ministry of Justice (see Box 6 below) developed its own standardised reporting template for certain types of evaluation. This can have additional benefits, as the department may provide guidance on how the results of an evaluation ought to be interpreted, and because templates enable readers to make comparisons between different interventions.

Recommendation

7. Follow reporting guidelines. Evaluation reports should follow established reporting guidelines to ensure analysis choices are explained completely and transparently.

How to apply this recommendation in government evaluation?

Evaluation reports should follow established reporting guidelines that are appropriate to the evaluation method used. Where it is proportionate to do so, departments should develop their own reporting guidelines which provide additional information over and above that detailed in existing standards.

6 For a comprehensive list of reporting guidelines, and other material to help improve research reporting, readers should consult the EQUATOR Network (Enhancing the Quality and Transparency Of Health Research).
7 (Yao et al., 2017).

Box 6: Reporting guidelines: Ministry of Justice, Justice Data Lab

The Ministry of Justice's *Justice Data Lab* uses a standardised reporting template to report on the results of all the interventions that it evaluates. This template provides clear information on the intervention of interest, a description of treatment and control groups, information on the outcome variables, and a comprehensive reporting of the results of the analysis.

The template also provides a consistent approach to describing effect sizes and uncertainty estimates, making it easier to compare the effects of different interventions. For example, effects are presented in terms of the number of people reoffending per 100 prisoners. Effect sizes are described so that they are understandable for a lay audience, and in numerical and visual form.

The template also includes a clear statement about what can – and what cannot – be inferred from a given analysis. This includes exemplar text on 'What you can say about the one-year reoffending rate'; and 'What you cannot say about the one-year reoffending rate' on the basis of the results presented in the evaluation report.

Communicating uncertainty

Communicating uncertainty is a core part of transparent reporting of quantitative evaluations. If uncertainty is not presented in a clear and accessible way, evaluation findings risk being misinterpreted as more conclusive and definitive than they truly are. Transparent communication of uncertainty requires going beyond presenting p-values, which are typically hard to interpret for non-technical audiences and are commonly misused and misinterpreted. In addition, these statistics are poorly suited to communicating how much uncertainty there is around a central point estimate, which is often the most relevant information for decision-makers.[8]

8 (Imbens 2021).

Recommendation

8. Communicate uncertainty. Evaluation reports should present measures of statistical uncertainty (for example, confidence intervals, standard errors) alongside point estimates and explain what these imply in accessible language.

How to apply this recommendation in government evaluation?

Evaluations using quantitative methods of evaluation should present measures of statistical uncertainty – such as confidence intervals, standard errors, or probability distributions – alongside point estimates. Such measures should show the range of plausible effects of an intervention and explain what those ranges mean in language accessible to non-technical audiences. Reports should avoid presenting uncertainty solely in terms of statistical significance.

Robustness and sensitivity analyses

Robustness and sensitivity analyses investigate whether the results from quantitative impact evaluations change when analysts make different choices in how the data is analysed. Many quantitative analyses have been shown to be sensitive to reasonable variation in analytical approaches.[9] It is therefore essentially to understand the extent to which evaluation findings are robust to different analysis strategies.

There are several ways of conducting robustness and sensitivity analyses for quantitative evaluations, ranging from less systematic to highly systematic approaches. In many cases, robustness checks are implemented in a relatively arbitrary manner, with several versions of the analysis being implemented where a few key analysis choices are varied (such as control variables, outlier handling, model specification and so on). Results corresponding to different choices are then presented, typically in a technical appendix to the evaluation report.

9 See, for example, (Breznau et al., 2022; Hoogeveen, Sarafogllou, Aczel et al., 2023; Hoogeveen, Sarafoglou, van Elk et al., 2023; Huntington-Klein et al., 2021; Schweinsberg et al., 2021; Silberzahn et al., 2018).

Helpful tools

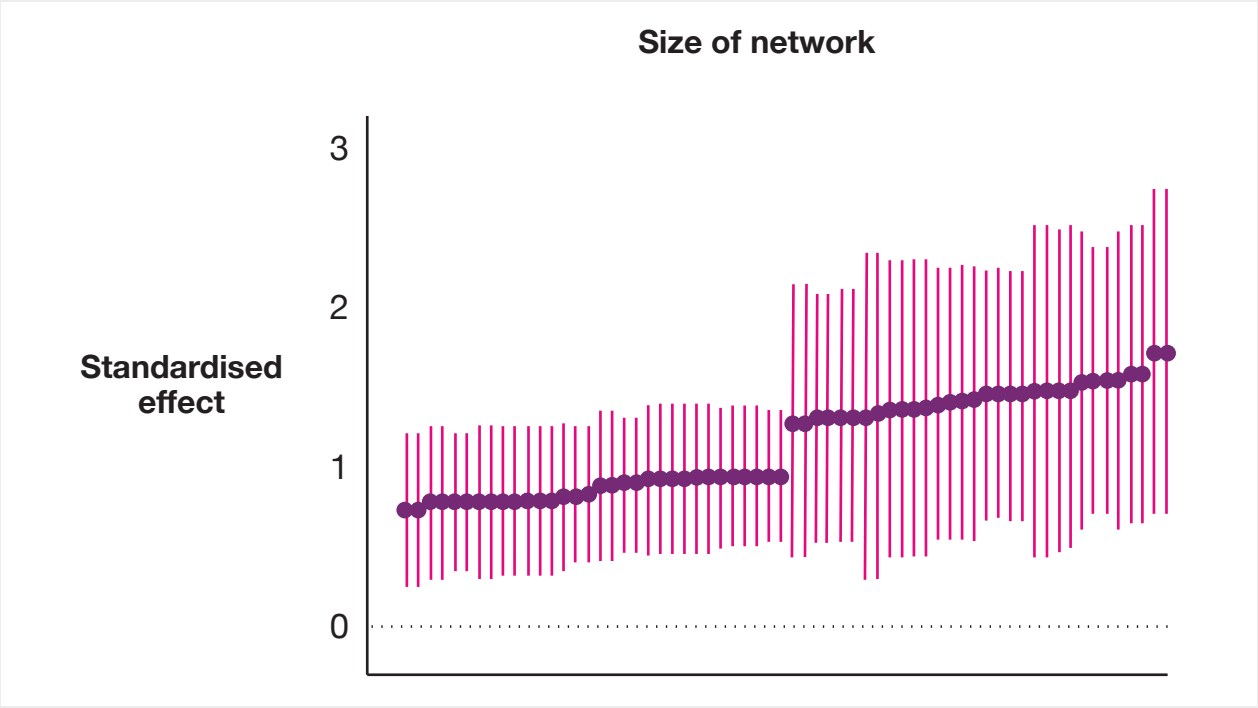
There are several more **systematic approaches** to robustness and sensitivity analysis.

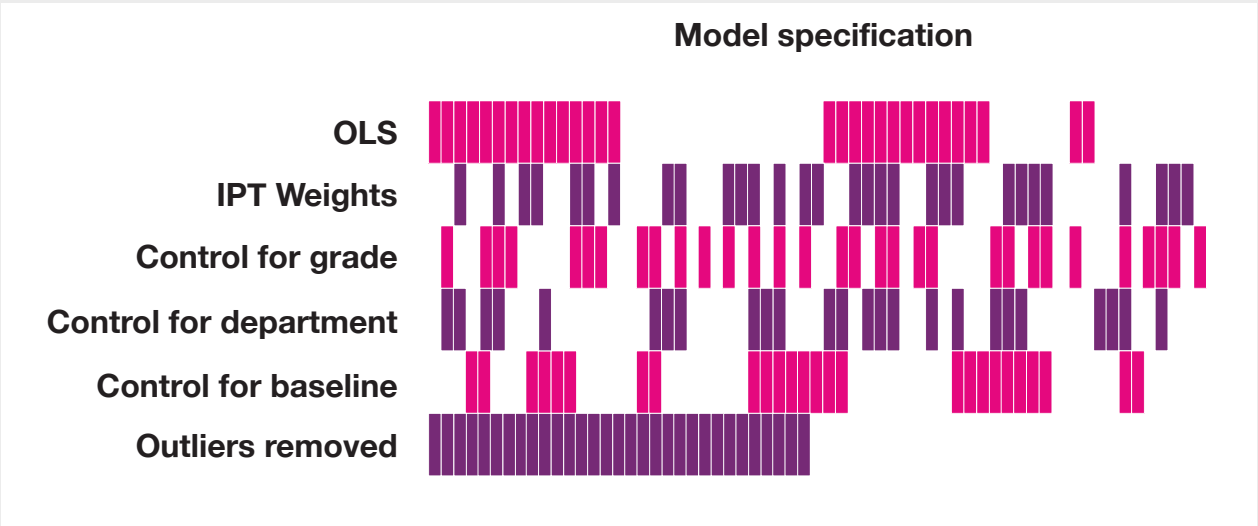
1. **Multiverse analysis.** A multiverse analysis is an approach that identifies all reasonable analysis choices and then systematically produces and reports results for every possible combination of these choices (Ganslmeier & Vlandas, 2025; Steegen et al., 2016). It is an alternative to reporting a single, potentially arbitrary, analysis accompanied by a small, arbitrary set of robustness checks.
2. **Specification curves** are a method of visualising the sensitivity of the results from a multiverse analysis (Simonsohn and others, 2020). A specification curve orders all estimates from the most negative effect of an intervention on an outcome, to the most positive effect of an intervention on an outcome. It condenses many robustness checks into a single, concise figure, making the full scope of uncertainty easy to grasp (see Box 7 for an example). Several tools are available to implement multiverse analyses and specification curves.[10]
3. **Many-analyst studies** involve multiple individuals or teams of analysts independently analysing the same data to produce an independent estimate of the effect of a policy intervention. The results of these independent analyses are then presented to demonstrate the robustness of the research to alternative analysis strategies. In contrast to a multiverse analysis, a many-analyst study limits the set of statistical analysis choices to those that an expert analyst would make. Recommended practice for many-analyst studies exists (Aczel et al., 2023). As the number of analysts required for such studies is usually quite large (in published many-analyst studies, the number of analysts varies between 14 (Huntington-Klein et al., 2021) and 161 (Breznau et al., 2022)), they are both time and resource intensive.

10 For instance, Boba (Liu et al., 2020) is a multiverse analysis package available in both R and Python, and the specr package provides specification curve plotting functions in R (Masur, 2020; Masur and Scharkow, 2020).

Box 7: Multiverse analyses: Evaluation of the Evaluation Academy

In an evaluation of the Evaluation Academy training programme, the Cabinet Office's Evaluation Task Force conducted a multiverse analysis to assess the sensitivity of their results to alternative analytical decisions. The researchers identified six sets of analytical choices that could have been consequential for the estimates presented in the report, including the control variables used, the type of regression model, whether outliers were removed, and whether survey weights were used. Treatment effects were then estimated for each combination of those choices. The results of the multiverse analysis for one outcome variable – the size of participants' networks in government – are summarised in a specification curve in the figure below.





The Y-axis of the top panel of the figure measures the estimated effect of participation in the Evaluation Academy on the *size of network* outcome, and the X-axis indicates the specification to which that estimate relates. The boxes aligned in the bottom panel indicate the analytical decisions behind each specification. For instance, the leftmost point estimate comes from an approach which used an Ordinary Least Squares (OLS) regression, which removed outliers, and which controlled for the grade of respondents.

A total of 64 analyses were conducted, and the specification curve summarises the results of these analyses. In this case, the effects are positive and significant regardless of which specification is used, though there is some variation in effect magnitude. The results demonstrate that whether outlier data points were removed from the analysis was the most consequential decision for the impact estimates from this evaluation.

Recommendation

9. Conduct and report sensitivity analyses. Analysts should conduct appropriate robustness checks, testing core findings against alternative analysis choices with results reported in the evaluation report.

How to apply this recommendation in government evaluation?

Analysts conducting quantitative evaluations should conduct appropriate robustness checks, testing core results against alternative analysis choices specifications. Results should be reported in the evaluation report, or in an annex to this report and summarised in the report. When proportional, analysts conducting quantitative evaluations should consider implementing more systematic approaches to robustness and sensitivity analyses, such as a multiverse or many-teams' analyses. If implementing such an analysis, this should be included in the statistical analysis plan.

References

Aczel, B., Szaszi, B., Nilsonne, G., Van Den Akker, O.R., Albers, C.J., Van Assen, M.A. *et al.* (2021) 'Consensus-based guidance for conducting and reporting multi-analyst studies', *eLife*, 10, e72185.

Arel-Bundock, V., Briggs, R.C., Doucouliagos, H., Mendoza Aviña, M. and Stanley, T.D. (2024) 'Quantitative political science research is greatly underpowered', *The Journal of Politics*. https://doi.org/10.1086/734279

Blanco, D., Schroter, S., Aldcroft, A., Moher, D., Boutron, I., Kirkham, J.J. and Cobo, E. (2020) 'Effect of an editorial intervention to improve the completeness of reporting of randomised trials: a randomised controlled trial', *BMJ Open*, 10(5), e036799.

Breznau, N. *et al.* (2022) 'Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty', *Proceedings of the National Academy of Sciences*, 119, e2203150119.

Brodeur, A., Cook, N.M., Hartley, J.S. and Heyes, A. (2024) 'Do pre-registration and pre-analysis plans reduce p-hacking and publication bias? Evidence from 15,992 test statistics and suggestions for improvement', *Journal of Political Economy Microeconomics*, 2(3), pp. 527–561.

C. Chang, A. and Li, P. (2022) 'Is economics research replicable? Sixty published papers from thirteen journals say "often not"', *Critical Finance Review*, 11(1), pp. 185–206. https://doi.org/10.1561/104.00000053

Chan, A.-W., Hróbjartsson, A., Haahr, M.T., Gøtzsche, P.C. and Altman, D.G. (2004) 'Empirical evidence for selective reporting of outcomes in randomised trials: comparison of protocols to published articles', *JAMA*, 291(20), pp. 2457–2465.

Cobo, E. *et al.* (2011) 'Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial', *BMJ*, 343.

Dudda, L. *et al.* (2025) 'Open science interventions to improve reproducibility and replicability of research: a scoping review', *Royal Society Open Science*, 12(4), 242057.

Ganslmeier, M. and Vlandas, T. (2025) 'Estimating the extent and sources of model uncertainty in political science', *Proceedings of the National Academy of Sciences*, 122(25), e2414926122. https://doi.org/10.1073/pnas.2414926122

Hardwicke, T.E. *et al.* (2018) 'Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition', *Royal Society Open Science*, 5(8), 180448.

Hoogeveen, S. *et al.* (2023a) 'A many-analysts approach to the relation between religiosity and well-being', *Religion, Brain & Behavior*, 13, pp. 237–283.

Hoogeveen, S. *et al.* (2023b) 'Many-analysts religion project: reflection and conclusion', *Religion, Brain & Behavior*, 13, pp. 356–363.

Huntington-Klein, N. *et al.* (2021) 'The influence of hidden researcher decisions in applied microeconomics', *Economic Inquiry*, 59, pp. 944–960.

Imbens, G.W. (2021) 'Statistical significance, p-values, and the reporting of uncertainty', *Journal of Economic Perspectives*, 35(3), pp. 157–174.

Ioannidis, J.P.A. (2005) 'Why most published research findings are false', *PLoS Medicine*, 2(8), e124.

John, L.K., Loewenstein, G. and Prelec, D. (2012) 'Measuring the prevalence of questionable research practices with incentives for truth telling', *Psychological Science*, 23(5), pp. 524–532.

Kerr, N.L. (1998) 'HARKing: hypothesizing after the results are known', *Personality and Social Psychology Review*, 2(3), pp. 196–217.

Liu, Y., Kale, A., Althoff, T. and Heer, J. (2020) 'Boba: authoring and visualizing multiverse analyses', *IEEE Transactions on Visualization and Computer Graphics*, 27, pp. 1753–1763.

Masur, P. (2020) 'How to do specification curve analyses in R: introducing "specr"', 2 January. Available at: https://philippmasur.de/2020/01/02/how-to-do-specification-curve-analyses-in-r-introducing-specr/ (Accessed: 20 August 2025).

Masur, P. and Scharkow, M. (2020) 'specr: visualizing specification curve analyses'. Available at: https://cran.r-project.org/web/packages/specr/vignettes/custom-plot.html

McKenzie, D. (2012) 'A pre-analysis plan checklist', *World Bank Blogs*. Available at: https://blogs.worldbank.org/en/impactevaluations/a-pre-analysis-plan-checklist

McConnell, B. and Vera-Hernandez, M. (2025) 'Going beyond simple sample size calculations: a practitioner's guide', *Fiscal Studies*, 46(3), pp. 323–348.

Nosek, B.A., Spies, J.A. and Motyl, M. (2012) 'Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability', *Perspectives on Psychological Science*, 7(6).

Obels, P., Lakens, D., Coles, N.A., Gottfried, J. and Green, S.A. (2020) 'Analysis of open data and computational reproducibility in registered reports in psychology', *Advances in Methods and Practices in Psychological Science*, 3(2), pp. 229–237. https://doi.org/10.1177/2515245920918872

Open Science Collaboration (2015) 'Estimating the reproducibility of psychological science', *Science*, 349(6251), aac4716.

Schweinsberg, M. *et al.* (2021) 'Same data, different conclusions: radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis', *Organizational Behavior and Human Decision Processes*, 165, pp. 228–249.

Silberzahn, R. *et al.* (2018) 'Many analysts, one data set: making transparent how variations in analytic choices affect results', *Advances in Methods and Practices in Psychological Science*, 1, pp. 337–356.

Simmons, J.P., Nelson, L.D. and Simonsohn, U. (2011) 'False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*, 22, pp. 1359–1366.

Simonsohn, U., Simmons, J.P. and Nelson, L.D. (2020) 'Specification curve analysis', *Nature Human Behaviour*, 4, pp. 1208–1214.

Steegen, S., Tuerlinckx, F., Gelman, A. and Vanpaemel, W. (2016) 'Increasing transparency through a multiverse analysis', *Perspectives on Psychological Science*, 11, pp. 702–712.

Stefan, A.M. and Schönbrodt, F.D. (2023) 'Big little lies: a compendium and simulation of p-hacking strategies', *Royal Society Open Science*, 10(2), 220346.

Wicherts, J.M., Veldkamp, C.L., Augusteijn, H.E., Bakker, M., van Aert, R.C. and van Assen, M.A. (2016) 'Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking', *Frontiers in Psychology*, 7, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Yao, X.I. *et al.* (2017) 'Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies', *JNCI: Journal of the National Cancer Institute*, 109(8).