

Methods note · Communities and governance

What a policy pilot can earn permission to scale

Sofia Mendes · 15 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

A pilot can establish delivery feasibility without resolving effectiveness, additionality, or performance at scale. This methods note develops a reading procedure for a hypothetical repair-voucher programme using the Magenta Book, the CONSORT pilot-trial extension, and Deaton and Cartwright's methodological analysis. It distinguishes invitation, enrolment, use, repair completion, and sustained service, and explains selection, intention-to-treat comparisons, missing data, uncertainty, and transportability. The fictional planning case contains no completed trial or measured outcomes. Scaling questions include repairer capacity, upfront-payment barriers, changing costs, and different participant populations. No policy saving or climate effect is estimated. The practical implication is to match progression decisions to the pilot's learning objective, pre-specified criteria, ethical constraints, and the supporting conditions a larger programme would require.

Keywords: policy pilots · feasibility · scaling · evaluation design

A source-based editorial text, not a report of original Still North research. Examples and imagery are illustrative.

First identify what the pilot was built to learn

A small policy programme can demonstrate that a booking system operates, that people will enrol, or that an intervention changes an outcome. These are different achievements. A pilot designed to learn whether delivery is feasible should not be read as a definitive effectiveness trial. Nor should an inconclusive small trial be treated as evidence that a policy has no useful effect.

This methods note offers a reading procedure for a proposed repair-voucher pilot. It draws on HM Treasury's Magenta Book, the CONSORT extension for randomised pilot and feasibility trials, and Deaton and Cartwright's methodological discussion of using trial evidence. It transfers principles explicitly rather than treating a medical reporting checklist as a legal requirement for local policy. The voucher scheme is hypothetical, and no pilot results, cost savings, or emissions effects are claimed.

The first page of a credible report should make its learning objective clear. Is it asking whether residents can use the process, whether repairers can deliver it, whether vouchers cause additional repairs, or whether the policy improves a wider outcome? Several questions may be legitimate, but the design and evidence must support each one separately.

Reconstruct the intervention and its alternative

Identify eligibility, offer size, duration, communication, repairer participation, payment timing, exclusions, and the treatment of unsuccessful repairs. A voucher is not only its face value. A scheme requiring an upfront payment and later reimbursement presents a different access barrier from one paid directly to the repairer. Those design details can influence both participation and outcomes.

Specify the baseline as well. Residents may already have repaired the product, used a warranty, visited a repair café, replaced it, or deferred action. The number of vouchers redeemed is not the number of additional repairs caused by the scheme. Some vouchers may subsidise repairs that would have happened anyway, while the offer may prompt enquiries or repairs even among people who do not redeem one.

The [Magenta Book](#) recommends building evaluation into design and examining process, impact, and value for money. This helps a reader connect operational records with a causal question. A press release reporting enthusiastic uptake may establish demand under the pilot's conditions while leaving additionality, durability, and distribution unresolved.

A hypothetical pilot and its honest questions

Imagine a proposed scheme inviting two hundred eligible households across two districts to use approved repairers. The numbers are invented planning assumptions, not a completed trial. The organisers want to know whether households can find a repairer, whether quotes are understandable, whether claims are processed reliably, and whether low-income residents face upfront-payment barriers. Those are useful feasibility questions.

Suppose the proposal also suggests randomly allocating the voucher offer among consenting eligible households. That could support an intention-to-treat comparison of outcomes between those offered the scheme and those not offered it, if conducted and analysed appropriately. It would estimate the effect of the offer under the study conditions, not automatically the effect of receiving a successful repair.

Comparing voucher users only with non-users would reintroduce selection. Users may have more repairable

items, more time, easier transport, or stronger motivation. Random allocation of the offer does not make subsequent voluntary use random. A reader should therefore check which groups appear in the analysis and whether exclusions were decided before the outcomes were known.

Two hundred invitations do not establish adequate statistical power. The required sample depends on the outcome, expected variation, clustering, uptake, and the size of effect relevant to the decision. If the pilot is intended primarily to estimate delivery parameters for a later evaluation, say so. A small study can be well designed for that purpose without pretending to resolve effectiveness.

Read denominators, missing data, and uncertainty

Follow participants through the process: eligible, contacted, enrolled, offered support, seeking a quote, receiving a repair, and available at follow-up. Each denominator answers a different question. A high completion proportion among people who reach the workshop may coexist with poor access among all eligible residents. Report reasons for non-participation where they can be collected ethically.

The CONSORT pilot extension emphasises feasibility objectives, participant flow, uncertainty, and pre-specified progression criteria. It also discourages treating pilot trials as underpowered definitive tests of effectiveness. For this policy example, its transferable lesson is to align the claims with the pilot's purpose and show what happened to the people initially considered.

Inspect the uncertainty around outcome differences, not only whether a threshold for statistical significance was crossed. A wide interval may include both a worthwhile effect and little benefit. That is a statement about what the design has resolved. It does not justify selecting the most favourable point estimate for a business case, or interpreting an uncertain estimate as exactly zero.

Ask what changes at scale

Even a credible local effect does not answer whether a larger programme will behave similarly. Pilot households may be volunteers reached through active community groups. Pilot repairers may have spare capacity and an unusual commitment to troubleshooting. A larger rollout can reach different residents, encounter less repairable products, and strain parts, staff, or appointment availability.

Deaton and Cartwright argue that using trial evidence in another setting requires knowledge beyond the trial estimate. The relevant application here is modest: compare the mechanisms and supporting conditions in the pilot and the proposed rollout. Randomisation strengthens a causal comparison within its design; it does not by itself establish that every district has the same repair market or access barriers.

Costs also change with scale. Initial software and training may be shared across more claims, but additional areas may require new staff, outreach, or transport support. A fully booked repair market may respond differently from one with spare appointments. Do not multiply a pilot's average cost by the national population or assume

every unit cost falls as volume grows. Identify which resources are fixed, expandable, or constrained.

Evaluate the outcome beyond redemption

A voucher programme may aim to maintain useful products, reduce replacement demand, improve affordability, or build repair capacity. Each outcome needs evidence. A completed repair is not necessarily durable service, and durable service is not automatically a permanently avoided purchase. Follow-up should establish whether products remained useful and what households did when a repair failed, without promising unrealistic lifelong tracking.

Climate claims require an additional comparison of repair inputs, product use, replacement timing, and relevant alternatives. A count of repairs cannot be multiplied by a generic product footprint without establishing displacement and consistent boundaries. The pilot can still justify operational improvements while its environmental effect remains unquantified. Keeping those conclusions separate protects useful learning from an unsupported headline.

Distribution deserves the same attention. Who could not afford a diagnostic fee, transport a heavy item, navigate an online form, or wait without an essential appliance? A scheme may be popular among users while excluding the people it was intended to help. Disaggregate only where sample size and privacy permit, and use qualitative accounts to understand mechanisms rather than assigning every subgroup a precise effect.

Decide what evidence earns progression

Agree progression criteria before reading the results: minimum operational reliability, acceptable safety and access arrangements, sufficient evidence about costs, and a clear plan for unresolved effectiveness questions. These criteria should reflect the decision and the seriousness of possible harms. They are not universal numerical thresholds to copy from another pilot. Record any later change and its reason.

Possible decisions include changing the offer, testing delivery in a contrasting area, commissioning a suitably designed impact evaluation, or proceeding with monitored expansion where the evidence and risk justify it. A phased rollout can preserve learning opportunities, provided allocation, fairness, capacity, and analysis are planned properly. Evaluation must not be used as a reason to withhold urgent safety obligations.

The limitation is that no short pilot can remove every uncertainty about sustained, system-wide change. The useful reading question is therefore what the pilot has earned permission to claim and what decision it can support. Scaling becomes more defensible when the next commitment matches the evidence, the remaining assumptions are explicit, and responsibility for learning continues after the announcement.

References

1. HM Treasury and Evaluation Task Force (2026). *Magenta Book: Central Government guidance on evaluation*.
2. Eldridge, S. M., Chan, C. L., Campbell, M. J., and colleagues (2016). CONSORT 2010 statement: exten-

sion to randomised pilot and feasibility trials. *BMJ*, 355, i5239.

3. Deaton, A., and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.