

Methods note · Everyday practice

After the click: testing whether evidence was understood

Kaisa Lehtinen · 16 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

Clicks and reading time describe exposure but do not establish comprehension, retention, or transfer. This methods note develops an evaluation for two hypothetical, accurate versions of a climate-evidence explainer. It draws on van der Bles and colleagues' uncertainty-communication experiments, Karpicke and Blunt's learning experiments, and the Magenta Book. Proposed tasks examine baselines, aggregate scale, and uncertainty while keeping trust, confidence, and policy agreement separate from learning. The analysis addresses audience selection, attrition, assessment-induced learning, delayed follow-up, accessible questions, and proportionate privacy. No reader data or communication effects are reported. Its practical implication is to choose evidence measures that test a defined capability and preserve disagreement over values, rather than treating platform engagement as proof that public understanding improved.

Keywords: public learning · evidence communication · comprehension · evaluation

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

Attention is an opportunity, not a learning outcome

A climate explainer can attract many visits without helping readers interpret evidence. It can also serve a small audience very well while looking unsuccessful on a traffic dashboard. Clicks, reading time, downloads, and shares describe parts of exposure and engagement. They do not directly show that a person understood a comparison, retained the explanation, or could use it when faced with a different claim.

This methods note develops an evaluation for two accurate versions of an evidence explainer. It draws on van der Bles and colleagues' original uncertainty-communication experiments, Karpicke and Blunt's learning experiments, and HM Treasury's Magenta Book. The synthesis concerns measurement design rather than a universal recipe for communication. Its proposed comparison and assessment tasks are hypothetical, with no reader data, learning gains, or communication effects reported.

The central question is what the communication is intended to make possible. Finding a source, recognising a baseline, interpreting a range, distinguishing a regional average from a local observation, and evaluating a policy option are different capabilities. A clear objective lets an editor choose evidence of learning rather than accepting whichever platform metric is easiest to obtain.

Keep distinct outcomes distinct

Reach asks whether intended audiences encountered the material. Comprehension asks whether they can explain its meaning. Retention asks what remains after time has passed. Transfer asks whether they can apply an idea in a new case. Trust concerns judgements about evidence or its source. Behaviour concerns subsequent action under real constraints. Each can matter, but none should stand in automatically for the others.

Van der Bles and colleagues' 2020 experiments distinguish perceived uncertainty, trust in numbers, and trust in a source. Their design is useful here because it measures specific responses to different communication formats rather than treating attention as the outcome. This note does not import an effect estimate from those experiments or assume that their audiences represent every public-learning setting.

Trust is not a substitute for understanding. A person may trust an institution while misreading its chart, or understand the chart while questioning a policy recommendation. Evaluation should not define learning as agreement with the communicator's preferred action. Evidence can inform a decision without deciding values, priorities, or the distribution of costs on the reader's behalf.

A hypothetical explainer comparison

Imagine two proposed versions of a short explainer about a regional temperature chart. Both retain the same data, source, baseline period, and uncertainty information. Version A explains the comparison in prose. Version B reorganises the explanation around an annotated sequence of questions. The hypothesis is that the second organisation helps readers interpret the chart, not that it makes the underlying evidence more certain.

Before publication, define a small set of tasks. Ask readers to identify the reference period, explain what a regional average can and cannot say about one place, and describe what an uncertainty range represents in this specific chart. Include an option to say that information is insufficient. A valid response should sometimes be a careful limit, not a confidently selected answer.

Then present an unfamiliar chart with different labels and ask a structurally similar question. This tests whether the reader can transfer the comparison principle rather

than repeat a sentence from the explainer. Keep the new task within the material's intended scope. Failure to solve an unrelated technical problem would say little about whether the communication met its learning objective.

These tasks are hypothetical design proposals, not a validated assessment instrument. A small accessibility and comprehension exercise should first check whether the wording, reading load, and response format introduce barriers. Ask people to explain how they interpreted a question. An ambiguous question can manufacture apparent misunderstanding even when the evidence explanation itself is clear.

Make the comparison fair

If the purpose is a causal comparison between versions, allocate exposure appropriately, hold relevant content constant, and specify the analysis before examining outcomes. Random assignment within a recruited sample can support an estimate for that study's conditions. It does not ensure that the sample represents people who never visit the website, read another language, or use assistive technologies.

A headline test on a live site may change which people click. Comparing comprehension only among those who subsequently volunteer for a quiz can then mix communication effects with audience selection. Identify whether the experiment concerns the headline, the explanation, or the complete route from invitation to learning. Track the relevant denominator and report attrition rather than silently excluding those who did not finish.

The Magenta Book distinguishes monitoring from impact evaluation and stresses evaluation design from the outset. Applied here, platform analytics can monitor reach and technical access, while a planned learning assessment addresses another question. An attractive dashboard is not an evaluation design merely because it contains several measures.

Assessment can itself produce learning

Karpicke and Blunt's 2011 experiments examine retrieval practice and conceptual learning using science texts and later assessments. Their relevance is not a universal effect size for climate explainers. It is the warning that asking someone to retrieve information is an activity that can change subsequent learning, rather than a perfectly passive measurement.

A pre-test may alert readers to the exact feature an editor later wants them to notice. An immediate quiz with feedback can teach the answer before a delayed assessment. Those features may be desirable in a learning programme, but they change the intervention being evaluated. Describe whether the intended product is an article alone or an article combined with questions and feedback.

Where delayed retention matters, pre-specify the interval and use an equivalent but not identical task where appropriate. Record who returns and whether return differs between versions. Readers who choose to complete a later assessment may be more interested or confident than those who disappear. A delayed result based on a selected minority should not be generalised to every person who clicked.

Score reasoning without rewarding confidence alone

Develop a simple rubric tied to the objective. For the regional-average task, credit recognising the aggregate scale and avoiding an unsupported local conclusion. For the range task, credit an interpretation consistent with the source's definition, including what the range does not represent. Where responses are open-ended, use examples to help assessors apply the rubric consistently without demanding identical wording.

Confidence can be recorded separately. A useful communication may increase justified confidence, reduce unjustified certainty, or reveal a previously unnoticed gap. Averaging confidence into a learning score can conceal those differences. Likewise, a person selecting the right multiple-choice answer by chance is not equivalent evidence to explaining why competing interpretations fail.

For policy-related tasks, score the handling of evidence rather than the policy preference. A reader may correctly describe an estimate and still favour a different trade-off because of access, distribution, or risk. Treating that disagreement as failed learning would turn evaluation into persuasion accounting. The communication's objective must not be quietly rewritten after the answers arrive.

Use clicks for the questions they can answer

Analytics remain useful for diagnosing discoverability, broken routes, device barriers, and whether an intended audience can find material. Low completion may indicate excessive length, inaccessible design, or an interruption unrelated to the content. Long reading time may indicate interest or difficulty. These interpretations need additional evidence, not an automatic rule that longer is better.

Collect learning data with informed participation, proportionate privacy protections, and a retention plan. People should not have to disclose sensitive beliefs or personal circumstances to access public evidence. Small organisations can use a modest, well-defined study instead of building a continuous surveillance system. The amount of data collected should follow the question and the risk, not the capacity of an analytics platform.

Limits and an editorial decision rule

A local assessment cannot establish lifelong learning, broad behavioural change, or public trust in general. Tasks can be imperfect, participants can be unrepresentative, and repeated testing can alter the outcome. Report uncertainty and differences between groups only where the evidence supports them. A version that performs better on one task may impose greater reading or access burdens elsewhere.

Choose a version because it supports the defined capability under acceptable access conditions, not because it maximises one convenient count. Report reach, comprehension, retention, and transfer separately where they were actually examined. The most useful evidence communication leaves a reader able to make a better distinction after the page is closed. Its evaluation should test that possibility directly rather than mistaking the opening of the page for its achievement.

References

1. van der Bles, A. M., van der Linden, S., Freeman, A. L. J., and Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, 117, 7672–7683.
2. Karpicke, J. D., and Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775.
3. HM Treasury and Evaluation Task Force (2026). *Magenta Book: Central Government guidance on evaluation*.