Journal of Systems Thinking

Reclaiming Science

Scientific Method 2.0

Derek Cabrera [a,@] and Laura Cabrera [a]
a: Cabrera Research Lab
@: dac66@cornell.edu

ABSTRACT: Who gets to say what counts as science? Universities train scientists, journals publish their work, funders choose questions, and companies sell products as scientifically proven. Their incentives can pull against the work that earns scientific authority. This paper develops Scientific Method 2.0, a systems account of how inquiry works and knowledge is earned. Reality consists of systems within systems. We study a part of that wider reality, build a model and test how well it fits. Scientific thinking is systems thinking. The method is the Love Reality Loop (LRL): a model generates questions and actions, reality returns information, and the model is evaluated and revised. The arbiter is reality. Method choices are guided by the Knowledge–Method Matching Matrix (KMMM), according to what is already known. The gold standard is convergence across abundant, diverse evidence. A computational experiment demonstrates the inquiry loop under specified conditions. Method choice also depends on feasibility, the value of further precision and the speed of useful feedback. Systems Random Controlled Testing over Time (S-RCT) describes how claims are tested across an accumulating research record. Different methods, variables and kinds of evidence can support the same explanation without being reduced to one data format. The scientific method belongs to people willing to practice its discipline. Reclaiming science means holding institutional authority accountable to that work and helping people understand, question and use it.

KEYWORDS: science | systems science | scientific thinking | Scientific Method 2.0 | Love Reality Loop | Knowledge–Method Matching Matrix | convergence | triangulation | S-RCT

---

1. Who gets to say what counts as science?

Universities train scientists. Journals publish their work. Funders decide which questions receive money. Companies put "scientifically proven" on products. Each exercises power over what the public recognizes as science. What makes that power legitimate?

Science earns its authority by asking good questions, producing evidence and correcting mistakes. Academia supports much of that work. It also manages careers, credentials and competition for money. What makes an institution successful can pull against what makes its science reliable.

The authority earned by inquiry can be coopted. Universities, journals, funders and companies borrow it through titles, affiliations, publications and claims of scientific proof. These may stand for substantial work. They can also carry a claim farther than its evidence warrants. We call that substitution *sciencewashing*.

The incentives are easy to see. A journal gains attention from a striking result. A researcher gains publications. An administrator protects a reputation. Each can gain something from a claim and lose something by admitting it is wrong. The test comes when correcting the error costs the institution something.

Experimental designs and randomized controlled trials (RCTs) show how this works. Doctoral students encounter them as the "gold standard" of evidence [1]. A design becomes a badge of scientific standing. Researchers receive extensive training in experimental design and much less in how to draw causal conclusions from many lines of evidence [2].

The pressure is familiar in business and policy schools. These schools study some of the messiest systems people have built, yet students learn to shape questions around the designs that leading journals reward. Advisors want their students published and hired. Training follows the incentive. Then the students return to labs lined with posters of science's poster children.

Look at those posters and you will find:

- Albert Einstein, who worked with 'thought experiments' and theoretical mathematics;
- Jane Goodall, who pioneered patient, unorthodox methods of observation;
- Murray Gell-Mann, who predicted a new particle with pencil and paper [3]; and
- Charles Darwin, who all but perfected the art and science of observation, and built his theory by synthesizing many lines of evidence.

None of them did their principal work with an RCT. Yet doctoral students are taught to fit their research to that mold. How did one design come to stand for science?

Habit. Researchers do what they were trained to do, and they were trained to do what gets published. When one design dominates training and rewards, more questions get shaped to fit it.

Strategic and ritual compliance. Authentic engagement in science means trying to find out what is true and being willing to change your mind. Academia can reward learning what an advisor, reviewer or funder expects and delivering it. Strategic compliance follows the rules to get a publication, grant or job. Ritual compliance follows them because that is how research is supposed to look. Both can produce the appearance of inquiry while the question and the evidence become secondary.

Medical envy. Education and psychology have borrowed medicine's evidence ladder to look more scientific. In 2013, England's Department for Education commissioned a physician to review evidence in schools. He pointed to medicine's gains from randomized trials [4]. The department followed with a call for RCTs in education and children's services [5]. In the United States, critics of the What Works Clearinghouse have questioned how useful its highest-rated evidence is to schools [6]. Even advocates recommend trials "where they are feasible" [4]. When funding and publication depend on a design, that qualification can get lost.

Manipulation of the public. "Clinically proven" can sell more certainty than a study earns. An analysis of 27 mathematics-curriculum RCTs meeting What Works Clearinghouse standards concluded that none

gave schools enough useful information to choose a curriculum; 12 had authors associated with the developers [6]. Others disputed parts of that analysis while agreeing that some studies were seriously flawed [7]. An Oxford team examined 431 performance claims for 104 sports products. Only 3 of the 74 studies it could appraise were of high quality and at low risk of bias [8]. One sports drink manufacturer supplied 101 trials from four decades, generally of poor quality and with very small effects [8]. A long list of experiments can make a persuasive advertisement without making a persuasive scientific case.

The choices come before the trial. Researchers and funders choose the question, population, comparison, outcome and setting. A well-run study can answer a question chosen to flatter a product while leaving the customer's question unanswered. No result has to be falsified. The distortion can lie in what was asked or what is claimed from the answer.

When people are taught to equate science with its institutions, an institutional failure can discredit sound evidence. A failure of marketing, governance or professional honesty is not necessarily a failure of the finding. People need to be able to tell the difference. Telling them to trust the institution more will not teach them how.

What counts as science is less settled than it looks. Dubova and Liu asked 92 scientists and 99 members of the US public to rate 18 classic descriptions of science, from a mirror of nature to detective work to Gell-Mann's search for beauty and elegance [9].[1] Scientists agreed with one another less than the public did. The public tended to see science as a reliable machine that produces objective truths by following well-defined procedures. Many treated creativity as a sign of pseudoscience. Scientists stressed the limits of method, the contingency of results and the role of improvisation. If working scientists do not share one picture of science, no single design can stand for it. The public's picture also sets science up to disappoint. When scientists disagree or revise a recommendation, people can read it as failure [9]. A public that expects certainty from every study needs to see where confidence comes from.

Reclaiming science requires standards people can use. This paper follows a series of questions: What makes thinking scientific? How does inquiry work? How do we choose a method? What makes a method better? How does evidence become knowledge we can trust? Each answer gives us what we need to ask the next question. First, what is science?

2. What makes thinking scientific?

You hear a strange noise in the car, consider possible causes and check when it happens. What the mechanic finds changes your explanation. A scientist makes this work more explicit and demanding. Both build models, ask questions and compare expectations with what happens. Einstein described science as a refinement of everyday thinking [10].

Science disciplines that thinking. Familiarity, confidence and a preferred explanation can all feel like knowledge. Science makes the questions, assumptions, methods and evidence open to examination. It asks what could show an explanation to be wrong and lets the answer matter. Anyone can practice that discipline. Doing it well requires care, knowledge and, for many questions, specialized training.

[1]The samples were not representative (mostly Santa Fe Institute visitors and Prolific users), but disagreement among any group of working scientists is enough to show that a shared understanding of science cannot be assumed.

2.1 The scientific method is not the academic method

The scientific method is not the academic method. By the academic method, we mean how people earn credentials, obtain appointments, win grants and publish work. These practices can support science. They can also reward a career while leaving the scientific question unanswered. An academic role tells us someone's place in an institution, not whether their thinking is scientific.

The scientific method belongs to people willing to practice it. That requires putting reality ahead of a preferred answer, ego, career, profit or institutional loyalty. We call this dispositional purity: curiosity, honesty and a willingness to be wrong. It shows in what a person does—seeking evidence that could challenge the model and changing it when the evidence requires. No appointment can confer that commitment, and no title excuses its absence.

This disposition takes practice. Sincere people can be mistaken, so expertise, careful methods and scrutiny still matter. The question is whether an investigator lets reality correct the model. Scientific authority is earned through that work.

2.2 Science is systems science

Reality consists of systems within systems. A cell belongs to an organism, an organism to an ecosystem, and that ecosystem to wider physical and social systems. The systems interact across levels. Whatever phenomenon we choose to study is part of something larger.

We cannot boil the ocean. We choose something we can study: this cell, this classroom, this market, this relationship. We draw a boundary, choose a scale and decide which interactions to examine. That is analytical isolation. Drawing the boundary does not sever the connections beyond it.

A study makes systems choices before collecting data. What is the thing being studied? What are its parts? How do they relate? From what perspective are we looking? A classroom study might focus on a student, a teacher, a lesson or a school. Each choice brings some relationships into view and leaves others outside the model. The findings have to be read with those choices in mind.

Scientific thinking is systems thinking because understanding a phenomenon requires organizing what we know about its parts, wholes, relationships and perspectives. DSRP names that organization: Distinctions identify things and what they are distinguished from; Systems organize parts and wholes; Relationships connect them; Perspectives establish the point and view from which they are considered [11]. The LRL puts that organization to work in inquiry.

Science studies systems within wider systems and builds models of how they work. Research itself works this way. Methods contribute different parts. Studies relate to one another. A field's knowledge depends on how its findings fit together. Scientific Method 2.0 makes those relationships explicit.

2.3 The practice of everyday science

Table 1 puts professional and everyday inquiry side by side. Both test mental models against reality and use the feedback to improve them. Use the same forms of inquiry, to the degree the question requires.

Action	A professional scientist…	An everyday scientist…

A Love of Science	A deep curiosity to know the truth even if it goes against one's beliefs...	A deep curiosity to know the truth even if it goes against one's beliefs...
Review what is known	Reviews the research literature.	Googles it, compares sources, and checks their reliability.
Observe	Observes a phenomenon systematically—sometimes for years, as Goodall did.	Pays closer attention. Watches and listens before jumping to conclusions.
Question	Conducts interviews or surveys.	Asks someone—or oneself—a clear question that does not plant the answer.
Experiment	Designs experiments to test possible explanations.	Tests assumptions using "mini-experiments" to see what happens...
Synthesize	Brings studies together to examine where their findings converge and diverge. Uses the findings to revise explanatory models.	Brings together multiple sources and diverse perspectives to arrive at a conclusion. Uses what holds up to improve the model.

Table 1. Scientific methods and their everyday equivalents. The same forms of inquiry help professional and everyday scientists test their mental models against reality and use what they learn to improve them. The question determines how much rigor, time, and effort the inquiry requires. Both begin with the same disposition: love reality; stay curious, even when it means being wrong.

3. How does scientific inquiry work?

What connects a model in someone's mind to the reality it describes? Textbook lists—observe, hypothesize, experiment, analyze, conclude—name useful activities. Working scientists move back and forth among them, changing questions and explanations as answers arrive.

The Love Reality Loop (LRL) describes that movement [11]. A mind builds a model. The model generates questions, predictions and actions that go toward reality through fitback. Information comes back through feedback. The mind compares the answers with its model and decides what to change. What it learns changes what it asks next.

Reality is constantly giving feedback. It does not reveal itself all at once or on demand. We put questions to it by watching, digging, measuring, intervening and seeing what happens. A question need not be a sentence, and an answer need not arrive in one neat piece. Digging asks whether fossils are present. Finding one can answer yes. Finding nothing in one hole may leave the wider question open. The answer depends on what the search could establish.

Information (I) is what a model holds about its subject. Organization (O) is how that information is distinguished, grouped, connected and viewed. The same information can mean something different when its organization changes. DSRP-484 names four patterns, eight paired elements and four dynamics. The pairs are identity–other, part–whole, action–reaction and point–view.

The dynamics describe how these work together. Equality concerns the relationship joining a pattern's two elements. Coimplication means that one element implies its partner: a part implies a whole. Simultaneity means that something can occupy several roles at once: a part can also be a whole, a relationship or a perspective. Enclosure means that each pattern requires the others. The formal account develops these

relationships [11]. Both Mind and Reality consist of organized information. IO means Information organized through O:

$$\mathbb{M} = \mathrm{IO} \qquad \mathbb{R} = \mathrm{IO} \qquad \mathrm{O} = \mathrm{DSRP\text{-}484}$$

The Mind's model ($\mathbb{M}$) sends questions, predictions, tests and actions toward Reality ($\mathbb{R}$) through fitback. Information returns through feedback, and the Mind evaluates and revises its model. Figure 1 places scientific inquiry within that loop:

$$\mathbb{M} \rightarrow \mathrm{Fitback} \rightarrow \mathbb{R} \rightarrow \mathrm{Feedback} \rightarrow \mathbb{M}$$

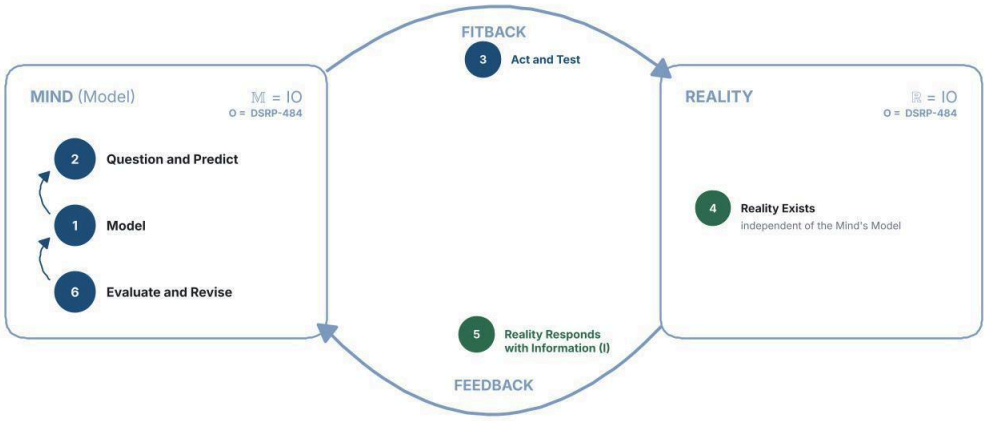


Figure 1. Scientific inquiry within the Love Reality Loop. The Mind models (1), questions and predicts (2), then acts and tests through fitback (3). Reality exists independently of that model (4) and responds with Information through feedback (5). The Mind evaluates the answers and revises its model (6). Adapted from [11].

Correspondence, C, measures how well the recovered organization matches Reality's declared reference structures. In the Zero-I test, success requires the structures to match one for one under the Mind's temporary names. Having the same number is not enough. Exact spellings are tested separately. Complete correspondence with the relevant reference organization is represented by 1. The limiting ideal is:

$$\mathrm{C}(\mathbb{M}_t, \mathbb{R}) \rightarrow 1$$

Here t marks successive encounters with Reality. Figure 2 shows why organization matters. Mind and Reality contain the same Things, a–d, but the Mind has captured only some of their grouping and relationships. Red marks Information; blue marks Organization. Inquiry can reveal what the model is missing. The arrow describes an ideal under suitable conditions, not automatic improvement. The Zero-I experiment tests whether the loop can reach that match in a declared computational Reality. [11]

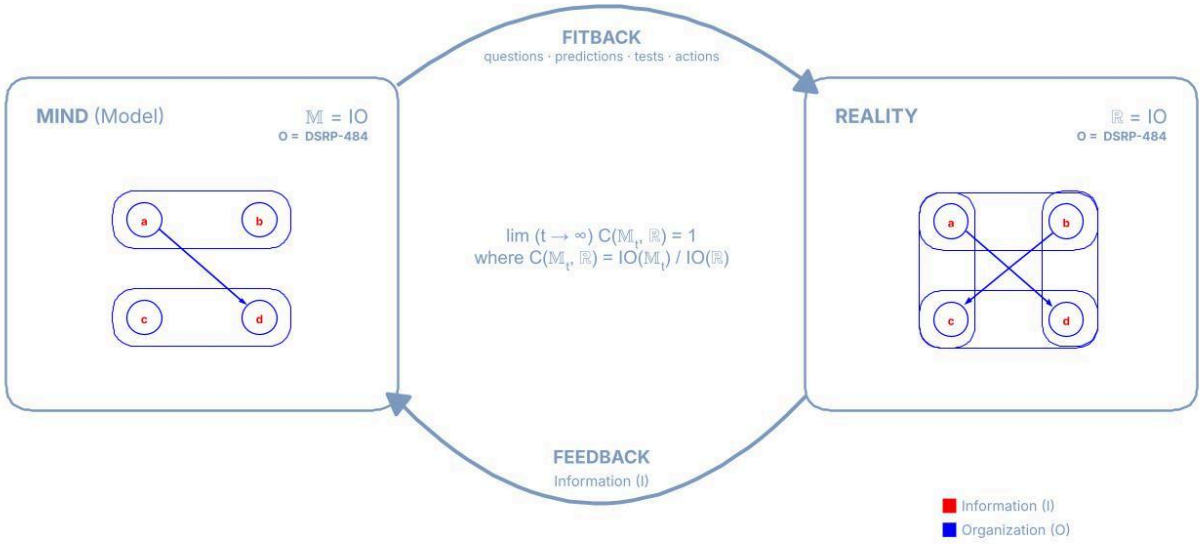


Figure 2. Model–Reality correspondence in the Love Reality Loop. Mind and Reality contain the same Things (a–d, red), organized differently (blue). Questions and actions go out through fitback; Information returns through feedback, allowing revision. The expression describes an ideal of structural correspondence. This is a schematic, not experimental data. Adapted from [11].

The loop is a description of inquiry that we can put to a test. Can a Mind discover a declared Reality's organization when it has to obtain its information by asking questions?

3.1 The Zero-I experiment

Could a Mind discover what Reality held without being handed the answer? We built two computational systems joined by two channels. The Mind received DSRP-484 and procedures for asking questions, but none of the Things, names or spellings it had to discover. Reality held reference structures declared before the runs. The Mind could not inspect them. [11]

The experiment uses three responses: yes, no and shrug. Yes affirms the queried condition. No negates it. Shrug leaves it indeterminate. A shrug is information, but it does not settle the question either way. In this apparatus, Reality shrugged when no applicable declared structure bore on the question. That did not establish absence. These discrete responses make the exchange executable; feedback in the world need not arrive in three neat units. [11]

The Mind asked whether another Thing was available. Yes let it distinguish and name that Thing. No ended the count. With the count channel unavailable, Reality shrugged and discovery stopped. No complete Thing, name or structure was transmitted. The Mind first reconstructed organization under its own names, then asked about spellings using a declared 94-symbol alphabet. The test compared its model with Reality's reference structures. [11]

3.2 What the data showed

Across 110 runs on six corpora—a cancer table, movie reviews, digit images, Mary Had a Little Lamb, the first chapter of the Tao Te Ching, and a panda sentence—the Mind recovered Reality's declared structures exactly in 110 of 110 runs. It recovered that organization before learning how Reality spelled the Things. [11]

The spelling phase recovered 7,434 of 7,514 Things exactly. The remaining 80 came from 20 cancer-table runs: four column names contained spaces, and the declared alphabet had no space. The Mind recovered their organization but could not recover a symbol it could not ask about. The failure exposed a limit of the available questions. [11]

The controls checked where the result came from. Switching off Reality's count channel stopped discovery. Starting without the Things required more questions than being given them, but the cost of organizing them was identical in 110 of 110 paired runs. Cost here counts questions opened by the Mind. Finding the Things cost their number plus one in every run: a question for each Thing and a final question establishing that there were no more. [11]

The experiment put inquiry on the bench. A Mind asked questions, received answers and built a model that could be checked against Reality. It demonstrated discovery within that apparatus: the reference structures were declared, answers were truthful and the response channel had no added noise. Natural inquiry also faces instrument error, incomplete access and conflicting reports. Correspondence can stall or worsen. The experiment establishes what happened under its stated conditions. [11]

The loop ran, and its result could be checked. Observation, questioning and testing offer different ways of obtaining and evaluating feedback. How should we choose among them?

4. How do we choose the right method?

Science is the work of developing and correcting knowledge through contact with reality. The scientific method is how that inquiry proceeds; the LRL makes the process explicit. Methods are the particular procedures we use, such as interviews, surveys, experiments and literature reviews. Methodology explains why those methods fit a question and how their findings work together.

Scientific Method 2.0 brings four relationships into one account. The method is the LRL. The arbiter is reality. Method choices are guided by KMMM. The gold standard is convergence. Table 2 shows how they fit. The version number names this formulation; people have long been asking reality questions and revising what they think.

Role	What fills it	What this means in practice
The method	Love Reality Loop (LRL)	A mind builds a model, asks and acts through fitback, receives information through feedback, then evaluates and revises the model.
The arbiter	Reality	A claim must hold up against what reality returns. A credential, institutional position or preferred conclusion cannot decide the answer.
Method choice	Knowledge–Method Matching Matrix (KMMM)	Match methods to the question and condition of knowledge. Consider what their feedback can establish, how soon it arrives and what it costs.
The gold standard	Convergence	Earn confidence through abundant, diverse evidence that survives scrutiny of alternatives and shared errors. S-RCT describes that cumulative testing over time.

Table 2. Scientific Method 2.0. The method, arbiter, basis for method choice and gold standard are distinct roles within one system of inquiry.

What we learn can change the question, expose an unclear construct or call for a different method. The Knowledge–Method Matching Matrix (KMMM) relates method choice to the condition of knowledge [12].

Figure 3 relates two continua. The vertical axis is the condition of knowledge: how much we know, and how well. It runs from nascent through developing to mature. The horizontal axis groups methods into observation, analysis and synthesis. The choice depends on what we know and what we need to find out.

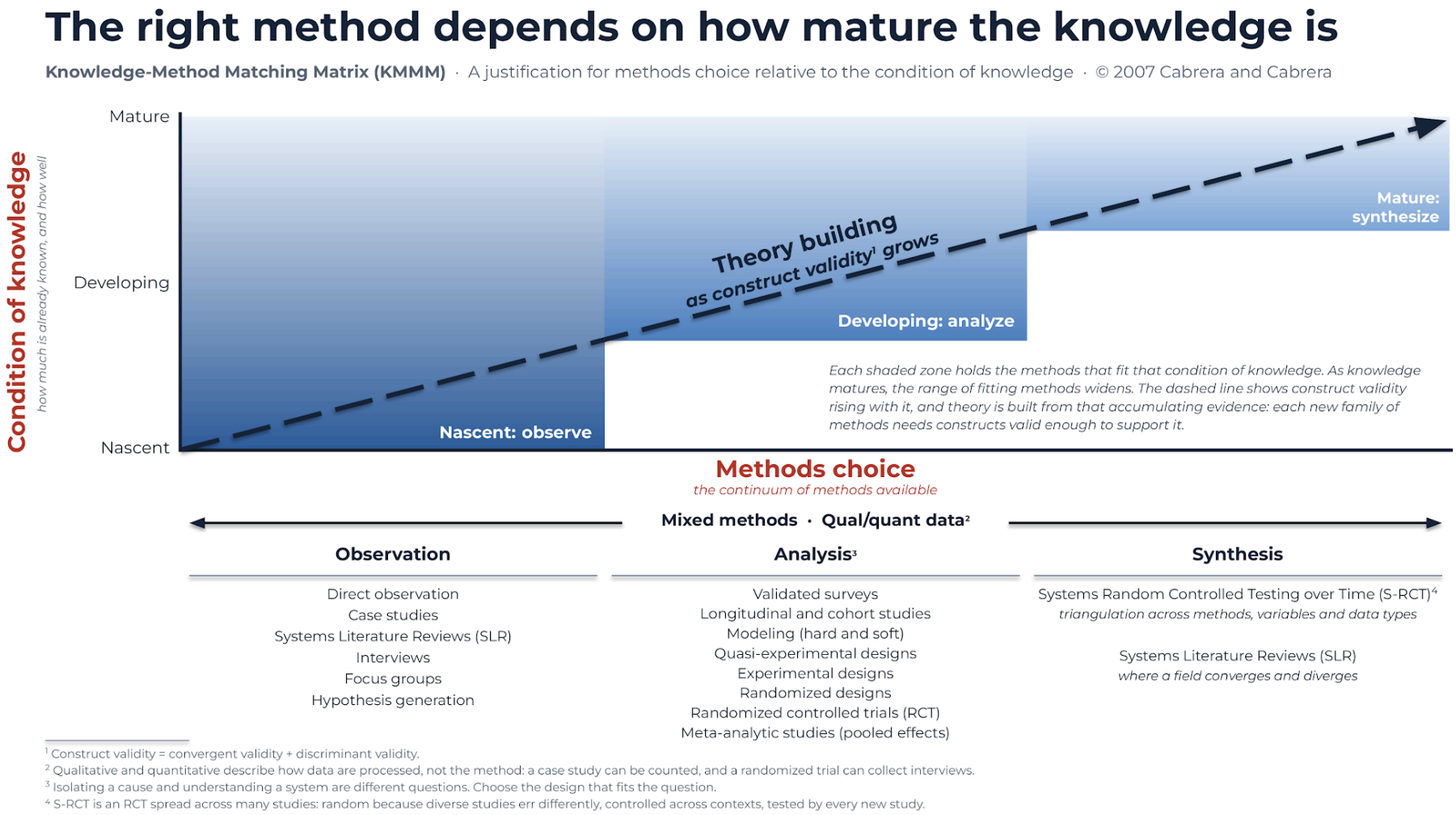


Figure 3. The Knowledge–Method Matching Matrix. Each shaded zone holds the methods that fit that condition of knowledge. As knowledge matures, the range of fitting methods widens, and earlier methods remain available. The dashed line shows theory building rising as construct validity grows. Mixed methods and qualitative and quantitative data span the whole continuum [13].

4.1 Every method is a test

All three families run the loop. Observation compares what we expect with what happens. Analysis takes findings apart to examine differences and relationships, including causal effects where the design allows. Synthesis puts findings together to see what explanation they support. Both taking apart and putting together can expose a mistake.

4.2 Nascent knowledge: observe

An experiment needs variables clear enough to manipulate or measure. A randomized trial also needs a defensible treatment, comparison and outcome. That understanding has to come from somewhere. If the constructs are unclear, a precisely executed design can give a precise answer to an unclear question.

Imagine an alien has just revealed itself on Earth. What would we manipulate? What would we measure? We do not yet know enough to choose. We would watch, perhaps from a distance. We might document encounters, ask the alien questions or interview people who met it. Each observation helps us build and revise a model. We develop hypotheses as we learn what needs explaining. The observation column includes direct observation, case studies, reviews, interviews, focus groups and hypothesis generation.

4.3 Developing knowledge: analyze

As constructs and variables become clearer, researchers can use validated surveys, longitudinal and cohort studies, models, quasi-experimental and experimental designs, randomized designs and RCTs. We might compare repeated encounters with the alien before attempting an intervention. Meta-analysis pools compatible quantitative estimates to examine effects and their variation. The matrix places pooled-effects work in analysis because it estimates specified relationships. It is also commonly called quantitative evidence synthesis.

Hard models, such as agent-based modeling (ABM) and system dynamics models, need well-specified variables and good data. Soft approaches, such as the Agent Based Approach (ABA) and causal loop diagrams, can work with less precisely specified variables [14]. The methods in the column are arranged by what their constructs require. Their order does not rank the strength of their evidence. Isolating a cause and understanding a system are different questions. Choose the design accordingly.

4.4 Mature knowledge: synthesize

As studies accumulate, we can ask where their findings converge, where they disagree and what remains open. The synthesis column includes systems literature reviews (SLR) and triangulation across methods, variables and data types. Section 6 explains how that record tests claims over time through S-RCT.

Observation remains useful as knowledge matures. So do the other earlier methods; the shading extends upward for that reason. Mixed methods can be used throughout. Qualitative and quantitative describe forms of data, not exclusive classes of method: a case study can count observations, and a randomized

trial can collect interviews. A field can also be mature in one construct and nascent in another. A study may therefore draw from several parts of the matrix.

4.5 Why construct validity matters

Some things are easier to identify and measure than others. Counting animals is easier than measuring empathy. The problem starts before statistics: what are we measuring?

Ask 100 people to count 37 sheep in a pen. Given a clear view and agreement about what counts as a sheep, we expect similar answers. Now ask them to count the empathy among 37 people in a room. What counts as a unit? What behavior would show it? Before comparing the counts, we need to know whether people are counting the same thing.

A construct is a concept we are trying to define and measure, such as empathy. Construct validity asks what a measure actually captures. Campbell and Fiske developed the multitrait–multimethod (MTMM) approach to examine it [15]. Different methods measuring the same trait should converge: convergent validity. Measures of different traits should distinguish them: discriminant validity. Empathy and niceness may be related, but an empathy measure needs evidence that it captures something beyond niceness. Observation, interviews and cases help clarify that distinction. Measurement and testing can then refine or challenge it. As this work develops, a field can make more precise claims and build better-supported theory.

Complexity makes this work harder. KMMM helps us see what a question still requires, including work on constructs that remain unclear. Then comes another question: can we do the study, and is what we would learn worth what it takes?

5. What makes a method better?

A method can be excellent under conditions we cannot create. Calling it "better" does not make those conditions available. We have to work with the people, time, data, access and ethical permissions we have. An infeasible method cannot answer the question in front of us. Its advantages on paper do not make it the better choice.

5.1 The F ≪ A problem is a general principle

Sampling makes the point. To eliminate sampling error, we could try to observe the entire population. But at what point would we learn too little more to justify the cost? A sample may answer the question well before we have counted everyone. We still have to justify what it tells us. The choice is a judgment about what knowledge we need, what uncertainty remains and what further precision would cost.

Science is about what is possible, not only what is ideal. An experimental design (ED) may offer stronger control than we can achieve in practice. A quasi-experimental design (QED) may answer the question under different assumptions, with less cost or delay. Those assumptions still need examination. The comparison includes what the designs can establish and what it takes to run them.

The same is true of modeling. ABM can represent detailed interactions when the variables, data and resources support a simulation. ABA lets us investigate systems where a detailed simulation is out of reach [14]. A model we cannot build supplies no evidence.

We express this recurring mismatch as:

$$F / A \ll 1$$

Here A is the number of questions to which an approach is relevant in principle. F is the number for which it is feasible. For many demanding approaches, F is far smaller than A. This is an expression of the gap between applicability and feasibility, not a measured ratio or a claim that every method has the same gap.

Data may be unavailable. Variables may be unclear. Cost, time, access or technical demands may put a design out of reach. Ethics may rule it out. Even where we could run the study, we should ask what the extra demand would buy. Overkill begins when the additional methodological demand contributes too little knowledge to justify what it consumes.

That judgment depends on KMMM. An early look at a subject may need enough feedback to expose a mistaken assumption or identify the next question. A faster comparison can be worth more than a more elaborate design whose answer arrives much later. A consequential causal claim may need the stronger design. The method should fit the knowledge needed next, including the quality, cost and speed of its feedback.

Fast-cycle feedback matters because inquiry is a loop. What could we learn through several useful cycles in the time required for one elaborate study? Each answer can improve the model and sharpen the next question. Sometimes added precision earns the delay. Sometimes the supposed ideal simply slows learning.

Speed and feasibility do not excuse an answer the evidence cannot support. A feasible design still has to address the question. If none can, the question remains open. Methodological judgment includes both what we can learn and whether that learning is worth the time and resources it takes.

5.2 Feasibility and scope are different limits

Even a well-run, feasible study has limits. Causes interact, effects feed back and interventions work differently in different settings [16]. An $X \rightarrow Y$ design examines specified effects under specified conditions. Experiments can include interactions and multiple outcomes, but every study selects part of the wider system.

An intervention may work only in combination with others. Tested alone, it can appear ineffective despite contributing to the solution. An effect in one setting may also fail to support a broad product or policy claim [16]. How far does the evidence reach?

Smoking and lung cancer show how causal knowledge can develop when randomized exposure would be unethical. Case-control and cohort studies, beginning with Doll and Hill [17], helped build the case. Different populations, designs and sources of evidence converged, and consistency across studies became one consideration in judging causation [18].

The case for S-RCT goes further than feasibility. Convergence across abundant, diverse evidence asks more of a claim about a wider system than any one successful trial can. The explanation has to hold up across different methods and conditions, disagreements, alternative explanations and possible errors.

Convergence is hard. An RCT that is feasible and well executed still contributes only part of that work. The next question is how the parts accumulate into knowledge we can trust.

6. How does evidence become knowledge we can trust?

Evidence earns its value through what it establishes. A careful observation does not have to wait for another investigator to become good evidence. Later work may confirm it, extend it, expose a limit or overturn it. The question is what that work adds. Different methods can reveal mistakes that one method would miss [2].

Goodall's observations did not become good because someone else later agreed. Their quality lay in the work itself. Corroboration can strengthen confidence without changing the quality of the original evidence. An independent study can be weak; the same investigator can produce a strong new test. Independence matters where it gives us another way to detect an error. It is not a badge of quality.

Consensus is agreement among people. Convergence is agreement among lines of evidence. Expert agreement matters when it reflects a sound reading of that evidence. It can also reflect shared assumptions, incentives or access to the same selected findings. We have to examine what supports the agreement.

Think of a crime scene with 20 witnesses and sources of evidence. They point toward Violet killing the professor in the library with the candlestick. One witness saw a struggle through the window. Another heard it from the hall. The forensics team found fingerprints. A camera recorded Violet entering the library. A timeline connects the events. Each contributes something different: presence, contact, opportunity or action. Some bear on only part of the explanation. Their combined force depends on how those parts fit and whether competing explanations can account for them.

The claim X is the hub; the spokes are the routes pointing toward it. They can involve different methods, variables, kinds of information, contexts, levels and times. They need not measure the same thing to support the same explanation. The investigator has to understand how they connect, what each establishes and where a connection might fail.

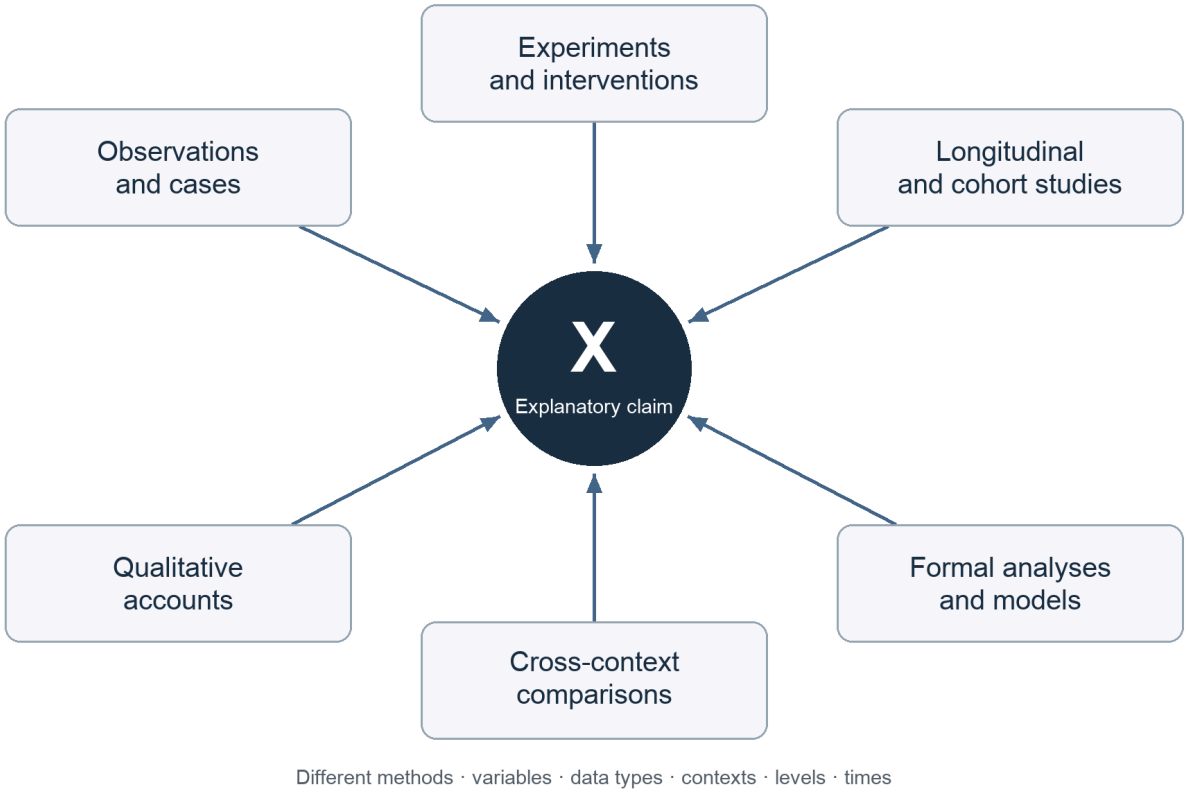


Figure 4. Convergence around a claim. Different evidential routes point toward X while retaining their methods, variables, data types, contexts, levels and times. DSRP supplies the common organization for relating them. The schematic illustrates a convergent case; spokes do not imply equal evidential weight or independent errors.

Now suppose all 20 witnesses watched through the same keyhole. Their agreement tells us more about what the keyhole shows. If something obstructed it, they could all be wrong together. Repetition and diversity contribute differently. The sports drink manufacturer's 101 trials likewise did not establish its case by their number alone [8]. What did each add? Which mistakes could they share?

6.1 Systems Random Controlled Testing over Time

We call this testing across studies Systems Random Controlled Testing over Time (S-RCT). It happens as a research record grows, often through work nobody planned together. Different routes test the same explanation. The evidence can remain in different forms. Random, controlled and testing describe what the record does as those routes accumulate [2].

Random. A wrong camera timestamp, an unreliable witness and a contaminated fingerprint make different mistakes. Agreement gains force when those mistakes cannot plausibly explain why the evidence points the same way. Studies may also share a measurement, assumption, funder or publication filter. We need to understand those connections. "Random" names the variation in errors across the record; it does not mean that investigators randomly assigned the studies or that errors automatically cancel.

Controlled. An explanation has more to answer for when evidence comes from different conditions. A finding that holds across places, populations, methods or scales becomes harder to explain as an artifact of

one condition. Where it fails can be just as informative: the difference may reveal a boundary or an interaction. Comparison across studies provides that control without requiring every study to manipulate the same cause.

Testing. Each new study can challenge the claim, narrow it or reveal another explanation. A study able to detect an earlier error adds something that repeating the same assumption cannot.

Over time. The LRL keeps running across investigators and studies. An answer can strengthen part of the explanation, reveal a condition or open a new question. Convergence is hard because the explanation has to hold together through those encounters. Table 3 compares this testing across a record with a trial's design.

	Randomized controlled trial (RCT)	**Systems Random Controlled Testing over Time (S-RCT)**
Random	Random assignment balances other factors in expectation	Different sources can expose different mistakes; examine what each adds and which errors are shared.
Controlled	A comparison group supports estimation of the treatment effect under the design's assumptions	Comparisons across contexts, populations and methods examine competing explanations and limits
Tested	Prespecified tests within a bounded study	Further studies can challenge, qualify or corroborate the claim
Scope	Specified effects and conditions within the trial's design	Relationships among findings across studies, settings and time
Works best when	Variables are well defined and a trial is feasible	A claim requires scrutiny across varied methods and conditions, and sufficient evidence has accumulated

Table 3. The RCT and the S-RCT. The comparison concerns the functions of designs and research records. Cross-study convergence requires attention to quality and shared bias; it does not reproduce randomized assignment or experimental control.

6.2 Bringing the research record together

Systems literature reviews show where findings recur, disagree or leave gaps. Triangulation brings approaches with different assumptions and weaknesses to bear on a question [2]. Campbell, Fiske and their colleagues developed this reasoning in social research [15] and [19]. Campbell's evolutionary epistemology described knowledge growing through variation and selection against reality [20]. S-RCT follows that testing across a research record.

Scientific Method 2.0 connects that work with the model–reality loop, method choice and feasibility. The computational experiment tests one part of the account. The wider argument depends on how the parts work together.

Meta-analysis pools compatible quantitative estimates to examine an effect or relationship and its variation. Studies can use different instruments or scales if a defensible transformation makes the results comparable. But some evidence cannot be pooled. The necessary measures or statistics may be missing, or the quantities may answer different questions. That is a feasibility limit. It does not make a meta-analysis inherently biased; it means the pooled evidence may cover only part of the case [21].

MTMM examines whether measures recover the same construct across methods and distinguish it from other constructs. Meta-analytic MTMM can examine those relationships across studies, provided the measurements and correlations support the comparison. Its contribution is evidence about construct validity [15].

An investigator cannot pool fingerprints, a recording and a witness's account as estimates of one effect. They can still converge. Evolution by natural selection likewise draws together findings about variation, inheritance, selection and change across generations. Different findings support different parts of the explanation [22]. They can fit together without starting in the same data format.

6.3 DSRP as the common organizational basis

Different information can share an organization. A fingerprint, testimony and a recording each involve distinctions, parts and wholes, relationships and perspectives. DSRP makes those explicit without erasing the differences. We can see which part of a claim a finding supports, how it connects to the explanation and the perspective from which it was produced. That is the common organization behind the hub and spokes in Figure 4.

DSRP has an accumulated evidential record. The current State of Systems Thinking report identifies 874 structural matches across 468 independently produced studies in 45 disciplines. Their authors studied their own subjects; later synthesis examined how the findings relate to DSRP. Other research examines how people use the structures and what changes when they learn them. Applications test their use elsewhere. Counterexample challenges try to break the inventory. Computation tests what it can do. Each asks something different of the same account [23,24].

Analyses of system dynamics, network theory and UML report representing 237 constructs without adding an organizational primitive: 60 from system dynamics, 35 from network theory and 142 UML candidates. These are formal containment analyses. They ask whether these established frameworks need any structural building block beyond DSRP [25].

The computational tests put the structures to work. cOde has represented the tested versions of BFO, CCO, DOLCE-Lite and PROV-O through its DSRP-derived structure. The roadmap reports 5,335 logical statements across fourteen ontology artifacts: BFO, eleven CCO modules, DOLCE-Lite and PROV-O. Holding the structures and reasoning with them are tested separately. The record includes joined BFO/CCO reasoning, cross-ontology conflict detection and transfer to held-out cases. Programs use the represented relationships while retaining differences among the sources. Each result names its corpus, reasoning task, supplied mappings and translation boundaries [26,27].

The challenge is open: find an organized structure DSRP cannot express, or one that requires another pattern, element or dynamic. The record contains 40 written-up challenges, all recorded as resolved. Some concern related puzzles. The challenges and resolutions are available for examination and further challenge [28].

We argue that this convergence makes DSRP the best-supported candidate for a universal ontology. The case is larger than the representation of another ontology. The same inventory holds up across independently developed formalisms, scientific findings, human studies, conceptual challenges and computation. Together, those results support its use here as the common organizational basis. Reality still decides whether a particular model fits.

Convergence develops across that work. No investigator has to plan or administer the whole record. A fingerprint can establish contact while leaving the time unknown. A model can explain how something happens without showing how often. Another finding can fill the gap or change the explanation. DSRP makes those relationships visible while keeping unanswered questions open. S-RCT describes how the case develops over time; it is not a prescribed sequence for producing agreement.

6.4 What have we earned: a hypothesis, a theory, a fact, proof?

A hypothesis is something we put to the test. “Violet did it” begins as a possibility. A theory connects findings into an explanation. As it survives tests and diverse evidence converges, it earns confidence. That is what we mean by theory being earned. A provisional theory can guide the next investigation; using it does not settle whether it is right.

The public often hears “theory” as “guess,” and that confusion has long been used against science [22]. Stephen Jay Gould described a scientific fact as something confirmed enough that it would be “perverse to withhold provisional assent” [22]. Facts need not be absolute certainties. They have survived enough testing that withholding assent would be unreasonable.

Life has changed and diversified: that is the fact of evolution. Theories explain how it happened [22]. The crime scene has the same distinction. Whether Violet killed the professor is one claim. How and why she did it require further explanation. Investigators can be more confident about one part than another. Debate about a mechanism does not erase the evidence for the phenomenon.

“Proven” exposes a problem of translation. A scientist can get a lump in the throat saying it: further evidence could always require revision. But the public often hears that reservation as uncertainty about the claim itself. What the scientist means as precision can sound unsure, evasive or noncommittal. Scientists have to speak into the public’s listening. They need to understand what people hear, as well as what they intend to say.

In ordinary conversation, “proven” means the evidence has settled the matter. When abundant, diverse research converges strongly enough to earn that conclusion, scientists should be willing to say so plainly. Evolution is a fact. Evolution by natural selection is supported by convergent evidence sufficient to place it beyond reasonable doubt–in other words, it's *proven* [gulp]. Questions about particular mechanisms, histories and circumstances remain open. Those questions do not make the established knowledge doubtful.

Scientific honesty requires communicating the strength of what we know as clearly as the limits. People should be able to tell what is established, what explains it and what remains open.

7. What would it mean to reclaim science?

Reclaiming science means holding scientific authority accountable to the work that earns it. Universities, journals, funders and companies should show how a claim was investigated, what the evidence supports and how it could be corrected. A title can tell us something about training or experience. It cannot establish whether a particular claim is true.

The same standard applies to universities, independent laboratories, companies and individual researchers. Working outside a university does not make a claim better or worse. The work has to stand up to examination.

Researchers should explain why their methods fit. What is the question? What is already known? What do the constructs and measures capture? What does the design assume, can it be done, and what could it establish? Readers should be able to follow the reasoning. A sophisticated technique cannot answer a question it was never equipped to address.

7.1 Change what institutions reward

Doctoral students should learn how methods work together as well as how to carry out particular designs. They need practice asking questions reality can answer, developing constructs and judging evidence across studies. Journals and funders should reward the observations, replications, contrary findings and syntheses that later tests depend on. Rewarding one design too narrowly spreads its blind spots across a field.

Public claims need the same care. Tell people what was studied, what was compared, what changed and how it fits with other evidence. A company should show the distance between the claim it tested and the claim it advertises. A university name or journal title cannot close that gap.

Put the public back in publication. People who support scientific work should be able to find its results, understand them and use them. Some ideas need technical language. Explain it. Keep the distinction the term makes, and give readers enough to follow the argument. Table 1 describes skills people can use themselves: compare sources, test expectations and change their minds when the evidence warrants it.

Institutions must protect the possibility of correction. Values can guide which questions we investigate. They cannot decide the empirical answers. Political, commercial or professional loyalty cannot shield a conclusion from examination. Feedback has to be allowed to change the model. An institution that decides which answers are acceptable before the inquiry blocks that part of the loop.

7.2 Science remains answerable to reality

Reality consists of systems within systems. We draw boundaries to study some part of it and build models of the parts, relationships and perspectives we find. Systems science makes that organization explicit. The LRL describes how a model meets reality, and the experiment shows the loop running under specified conditions.

The method is the LRL. The arbiter is reality. KMMM guides method choice according to what we know and need to learn next. Feasibility includes the cost and speed of useful feedback. Convergence is the gold standard: an explanation holding up across abundant, diverse evidence over time.

The scientific method is not owned by academic roles. People practice it through a disposition that lets reality correct what they think, even when the answer costs them something. Institutions earn scientific authority by supporting that work and keeping it open to scrutiny. Reclaiming science means holding the power to define, fund, publish and sell it accountable to the evidence, and teaching people how to investigate and evaluate claims themselves. Science answers to reality. Its benefits belong to people.

References

1. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG: An International Journal of Obstetrics & Gynaecology. 2018;125: 1716.

2. Munafò MR, Davey Smith G. Robust research needs many lines of evidence. Nature. 2018;553: 399–401.

3. Johnson G. Physicists recover from a summer's particle "hangover." 2016. Available: https://www.nytimes.com/2016/10/18/science/physics-particles-cern.html

4. Goldacre B. Building evidence into education. London: Department for Education; 2013.

5. Department for Education. New randomised controlled trials will drive forward evidence-based research. 2013. Available: https://www.gov.uk/government/news/new-randomised-controlled-trials-will-drive-forward-evidence-based-research

6. Ginsburg A, Smith MS. Do randomized controlled trials meet the "gold standard"? A study of the usefulness of RCTs in the What Works Clearinghouse. American Enterprise Institute; 2016.

7. Slavin R. Is all that glitters gold-standard? 2016. Available: https://www.huffingtonpost.com/entry/is-all-that-glitters-gold_b_9584118.html

8. Heneghan C, Howick J, O'Neill B, Gill PJ, Lasserson DS, Cohen D, et al. The evidence underpinning sports performance products: A systematic assessment. BMJ Open. 2012;2: e001702.

9. Dubova M, Liu QE. Everyone knows what science is…until you ask. Royal Society Open Science. 2026;13: 260933.

10. Einstein A. Physics and reality. Journal of the Franklin Institute. 1936;221: 349–382.

11. Cabrera D, Cabrera L. The Love Reality Loop: A unified organizational theory of mind and reality. Journal of Systems Thinking. 2026;6.

12. Cabrera D, Cabrera L. The Knowledge-Method Matching Matrix. Evaluation Systems and Systems Evaluation: Building Capacity and Tools for Enhancing STEM Education Evaluation. 2006.

13. Cabrera D, Cabrera L. Knowledge–Method Matching Matrix. 2026.

14. Mulyono YO, Sukhbaatar U, Cabrera D. “Hard” and “soft” methods in complex adaptive systems (CAS): Agent based modeling (ABM) and the agent based approach (ABA). Journal of Systems Thinking. 2023;3. doi:10.54120/jost.000009

15. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 1959;56: 81–105.

16. Cabrera D, Cabrera L. From one cause to webs of causality: Empirical evidence for a cognitive framework to solve wicked problems. 2025.

17. Doll R, Hill AB. Smoking and carcinoma of the lung. British Medical Journal. 1950;2: 739–748.

18. Hill AB. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine. 1965;58: 295–300.

19. Webb EJ, Campbell DT, Schwartz RD, Sechrest L. Unobtrusive measures: Nonreactive research in the social sciences. Rand McNally; 1966.

20. Campbell DT. Evolutionary epistemology. In: Schilpp PA, editor. The philosophy of Karl Popper. Open Court; 1974. pp. 413–463.

21. Higgins JP, Li T, Deeks JJ, editors. Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Cochrane; 2024. Updated August 2023. Accessed November 11, 2026.

22. Gould SJ. Evolution as fact and theory. Discover. 1981;2: 34–37.

23. The State of Systems Thinking Report. 2027 conference edition. Working report. 2026.

24. DSRP Evidence Meta Synthesis v1.2: Convergence across five bands of evidence. Research synthesis. November 4, 2026.

25. Bibliography Track Synthesis v2.1. Systems literature review of the DSRP bibliography. November 4, 2026.

26. cOde Executable Band Synthesis v1.1. Research records through November 2, 2026. November 4, 2026.

27. cOde Project Roadmap. Ontology of Ontologies, ontology tests and evidence records. Research record updated November 2, 2026.

28. Counterexamples and Resolutions Track Synthesis v1.1. Systems literature review of 40 counterexample cases and 17 resolutions. November 4, 2026.