

A SCIENTIFIC AND INTELLECTUAL HISTORY

# History of Diagnostics, Observability, and Debugging in Social Sciences

Social evidence and the examination of explanations

— *Seven parts and 35 chapters* —

**Concept by Dmitry Vostokov, [DumpAnalysis.org](https://DumpAnalysis.org)**

Developed with AI assistance (GPT-6 Astra Max and Fable 5.1 Extra) and checked against the sources listed in this article.

*Version 4 | September 2026*

Historical synthesis with 88 references, chronology, glossary, and evidence taxonomy

## Abstract

Social sciences investigate systems whose participants interpret situations, respond to measurements, and sometimes change their behavior because an investigation is taking place. Their evidence includes census returns, questionnaires, field notes, interviews, administrative records, maps, experiments, and digital traces. In this history, we examine how such artifacts became available for analysis, how researchers used them to distinguish competing explanations, and how errors in evidence and interpretation were discovered and corrected.

We follow developments in sociology, anthropology, economics, political science, psychology, demography, human geography, education, social work, and organizational research. Early criticism of historical reports and political arithmetic is followed by social surveys, statistical sampling, construct validation, causal inference, institutional analysis, and computational social science. Several chapters concern practices that explicitly used diagnostic language. Elsewhere, diagnostics and debugging provide our present-day interpretation of an earlier activity.

The main distinction is between a social condition, its recorded representation, and the explanation constructed from that representation. A change in a recorded rate can arise from a change in behavior, a new reporting rule, a different population, or an error in processing. Correcting the calculation resolves only one of these possibilities. We therefore examine the acquisition of evidence together with the assumptions required to interpret it.

Seven parts and 35 chapters are followed by a selected chronology, glossary, and evidence taxonomy. Connections to our pattern-oriented software diagnostics concern recurring analysis problems and relationships between artifacts. They help organize an investigation while preserving the differences between a software execution, a social interaction, and a historical explanation.

## How to read this history

We use diagnostics for the investigation of a problematic condition, result, or explanation. In social work and institutional research, this vocabulary has its own documented history. In other cases, we use it as an analytical connection. We do not suggest that every survey researcher, ethnographer, or economist understood their work as diagnostics.

Observability has a precise meaning when a specified dynamical model allows internal states to be reconstructed from outputs. Most historical social investigations do not establish that property. For them, we distinguish the visibility of records, the quality of measurements, the identifiability of an explanation, and access to participants' interpretations. More visible activity does not necessarily make a social process identifiable.

Debugging is literal when we correct research software, a coding rule, or a spreadsheet formula. Applied to a questionnaire, classification, explanatory model, or institutional arrangement, it is a present-day analogy for locating and correcting a specified defect. People and communities are not defective

components. Disagreement about an institution's purpose cannot be resolved by treating one preferred purpose as an unquestionable specification.

The chapters follow an approximate historical progression within thematic parts. Methods overlap in time and continue to coexist. Qualitative observation did not disappear with probability sampling, and digital traces did not remove the need for interviews. The selected cases include contributions outside the conventional European and North American sequence, but the book is not an exhaustive history of every regional tradition.

Square brackets identify references. Publication years are distinguished from the dates of translations, reissues, and digitization. Examples introduced as hypothetical are explanatory constructions. Connections to our analysis patterns are interpretations and proposed applications, not claims that earlier authors used the same pattern language.

### Neighboring practices and their questions

| Practice                          | Primary question                                                      | A limit to preserve                                              |
|-----------------------------------|-----------------------------------------------------------------------|------------------------------------------------------------------|
| <b>Description and monitoring</b> | What was recorded, where, and when?                                   | A recorded distribution does not explain its origin.             |
| <b>Measurement and validation</b> | What does this indicator support us in saying?                        | A repeatable score can measure the wrong construct.              |
| <b>Causal investigation</b>       | What difference would a specified intervention make?                  | A comparison requires assumptions about the missing alternative. |
| <b>Interpretation</b>             | What did an action or statement mean in its setting?                  | An analyst's category may differ from a participant's category.  |
| <b>Evaluation and correction</b>  | What failed relative to an explicit purpose, and what can be changed? | Purposes, effects, and authority may be contested.               |

# Contents

Seven parts and 35 chapters

| SECTION                                                                     | PAGE      |
|-----------------------------------------------------------------------------|-----------|
| <b>PART I · SOCIAL RECORDS AND THE FORMATION OF DIAGNOSTIC INQUIRY</b>      | <b>6</b>  |
| 1. Vocabulary and historical scope                                          | 6         |
| 2. Historical reports and the plausibility of social explanations           | 7         |
| 3. Political arithmetic and population evidence                             | 7         |
| 4. Moral statistics and the problem of normality                            | 8         |
| 5. Social surveys, reform, and case diagnosis                               | 9         |
| <b>PART II · EXPLANATION, OBSERVATION, AND THE CONSTRUCTION OF MEASURES</b> | <b>11</b> |
| 6. Social facts, meanings, and structural explanations                      | 11        |
| 7. Ethnography and the preservation of context                              | 12        |
| 8. Probability sampling and the diagnosis of survey error                   | 13        |
| 9. Psychological measurement and construct validation                       | 14        |
| 10. Comparability across groups, languages, and settings                    | 15        |
| <b>PART III · EXPERIMENTS, CAUSAL INFERENCE, AND HISTORICAL COMPARISON</b>  | <b>17</b> |
| 11. Experimental design and rival explanations                              | 17        |
| 12. Econometrics and the identification of a model                          | 18        |
| 13. Potential outcomes, causal graphs, and observational comparisons        | 19        |
| 14. Quasi-experiments and naturally occurring comparisons                   | 20        |
| 15. Comparative cases and process tracing                                   | 20        |
| <b>PART IV · INDICATORS, SEQUENCES, NETWORKS, AND ORGANIZATIONS</b>         | <b>22</b> |
| 16. National accounts and social indicators                                 | 22        |
| 17. Longitudinal studies and histories of transition                        | 23        |
| 18. Sociometry and social network evidence                                  | 24        |
| 19. Spatial evidence, aggregation, and ecological inference                 | 24        |
| 20. Interaction records and organizational failure                          | 25        |

| SECTION                                                                            | PAGE      |
|------------------------------------------------------------------------------------|-----------|
| <b>PART V · INTERVENTION, INSTITUTIONAL DIAGNOSIS, AND REFLEXIVITY</b>             | <b>27</b> |
| 21. Program evaluation and the explanation of outcomes                             | 27        |
| 22. Action research and participation in the investigation                         | 28        |
| 23. Institutions and diagnosis beyond universal remedies                           | 29        |
| 24. Reflexivity, predictions, and performance targets                              | 29        |
| 25. Authority, ethics, and the ownership of evidence                               | 30        |
| <b>PART VI · COMPUTATIONAL EVIDENCE AND THE EXAMINATION OF RESEARCH</b>            | <b>32</b> |
| 26. Digital traces and computational social science                                | 32        |
| 27. Simulation and the distinction between generation and explanation              | 33        |
| 28. Residuals, missing data, and sensitivity to analytical choices                 | 34        |
| 29. Replication, preregistration, and reproducible analysis                        | 35        |
| 30. Algorithms and generative models as social research instruments                | 36        |
| <b>PART VII · DIAGNOSTIC CASES, PATTERNS, AND THE PRESERVATION OF EXPLANATIONS</b> | <b>37</b> |
| 31. Polling failure and the reconstruction of an analysis                          | 37        |
| 32. Mixed evidence and the investigation of disagreement                           | 38        |
| 33. Connections to pattern-oriented software diagnostics                           | 38        |
| 34. A worked example of a misleading improvement                                   | 40        |
| 35. Diagnostic memory and the limits of reconstruction                             | 41        |
| <b>CONCLUSION</b>                                                                  | <b>42</b> |
| <b>APPENDIX A. SELECTED CHRONOLOGY</b>                                             | <b>43</b> |
| <b>APPENDIX B. GLOSSARY</b>                                                        | <b>46</b> |
| <b>APPENDIX C. EVIDENCE TAXONOMY AND BLIND SPOTS</b>                               | <b>49</b> |
| <b>REFERENCES</b>                                                                  | <b>51</b> |

**PART I****Social records and the formation of diagnostic inquiry****1. Vocabulary and historical scope**

Let us first consider what a social investigation may need to explain. A school records increasing absence. A neighborhood reports fewer crimes. A survey suggests declining trust. Each statement describes a relationship between a social activity and an observation procedure. Before explaining the activity, we need to establish what the procedure recorded and which population the statement concerns.

The same apparent problem can exist at several levels. A low participation rate may reflect inaccessible services. It may also reflect a restrictive definition of participation, an incomplete register, or an incorrect denominator. These possibilities lead to different investigations. Recalculating a percentage is appropriate for a denominator error. It cannot determine whether the institution offers a service that people need.

There are also historical uses of diagnostic language that do not depend on a software analogy. Mary Richmond's *Social Diagnosis*, published in 1917, organized the assessment of social evidence for casework. Much later, Elinor Ostrom's 2007 diagnostic approach examined how the characteristics of social and ecological systems affect institutional analysis. These are distinct projects, with different units of inquiry and purposes. Their shared vocabulary does not establish an uninterrupted school of social diagnostics. [1, 2]

For our purposes, a diagnostic artifact is a preserved representation that can support or challenge an explanation. A completed questionnaire is one such artifact. So are an interview recording, a case file, a table of transitions, and the code that converts administrative records into a published result. Their evidential value depends on how they were produced and on the question being asked.

We can also distinguish a problematic condition from a problematic description. Suppose a service's waiting time appears to improve after its reporting system changes. The improvement may be real, or it may result from recording the first response instead of completion. A correct description of the new measure can still be an incorrect answer to the old question. Here the diagnostic boundary lies between the institution's activity and its representation.

The history developed in this book concerns the creation and examination of these boundaries. Social sciences made populations countable, meanings discussable, comparisons more explicit, and research operations reproducible. They also exposed limits to these achievements. A person may refuse an interview for reasons relevant to the inquiry. An institution may respond to a performance target. An explanatory category may become part of the social world it describes. We need to carry these relationships into the analysis rather than remove them as inconvenient exceptions.

## 2. Historical reports and the plausibility of social explanations

Written records preserve selected activities long before a discipline called sociology exists. Registers, accounts, chronicles, legal proceedings, and correspondence allow later readers to reconstruct relationships that are no longer directly observable. However, the preservation of a statement does not establish its accuracy. We have to examine who produced it, which conventions governed its production, and why this particular record survived.

Ibn Khaldun's *Muqaddimah*, drafted in 1377, provides an important example of examining reports through knowledge of social organization. He discussed sources of error in transmitted history and the need to assess whether reported events were compatible with the conditions of human association. Franz Rosenthal's complete English translation appeared in 1958. The medieval work and its modern translation are different historical artifacts, and their dates should remain separate. [3]

The diagnostic connection is the examination of a report against constraints. A claim about the size of an army, for example, raises questions about recruitment, supplies, and the population from which it could have been drawn. A plausible narrative may fail such a check. Conversely, an unusual event should not be rejected simply because it does not fit an analyst's preferred generalization. A constraint is useful only when its own basis can be examined.

Source criticism also concerns dependence between reports. Five chronicles may repeat one earlier account. Their agreement then preserves a transmission history rather than five independent observations. In modern terms, we need the provenance of the evidence. The number of surviving statements can greatly exceed the number of independent encounters with the event.

Absence introduces another difficulty. A missing administrative entry may mean that nothing happened, that a clerk omitted the entry, that the relevant activity was outside the register's scope, or that a page was lost. We cannot choose among these explanations from the gap alone. Comparison with other records may help establish what normally generated an entry and whether that process was operating at the time.

This problem connects historical investigation with trace analysis without making the sources equivalent. A chronicle has an author, audience, and narrative purpose. Its omissions cannot be interpreted as if they were omissions from an instrument with a fully specified logging policy. We can still ask a common question: what relationship connects an event with the artifact that represents it? Establishing that relationship is part of diagnosis.

## 3. Political arithmetic and population evidence

John Graunt's 1662 *Natural and Political Observations* examined London's Bills of Mortality. The importance of the work for this history lies in the movement from individual entries and recurring lists to comparisons of population patterns. A record originally produced for one administrative and public

purpose became material for another form of inquiry. William Petty's *Political Arithmetick*, published posthumously in 1690, likewise treated numerical estimates of people, resources, and economic activity as material for political reasoning. [4, 5]

Counting creates a new diagnostic possibility. We can compare periods, distinguish a local fluctuation from a broader change, and ask whether a proposed explanation is consistent with several distributions. The comparison remains dependent on the recording system. A mortality category is also a classification made under particular knowledge and institutional arrangements. A change in its use may resemble a change in the underlying condition.

A count also requires a population at risk. Ten recorded events in a small population and ten in a large population do not have the same interpretation. Where population estimates are uncertain, the uncertainty extends into the rates calculated from them. Adding decimal places to the rate does not improve the underlying population estimate.

Consider a hypothetical register in which a new office begins recording events previously handled informally. The resulting increase is evidence of a changed administrative boundary. It may coexist with an unchanged rate of the activity itself. The useful diagnostic comparison includes neighboring offices, earlier definitions, and records of the change in procedure. A graph of the count alone leaves the explanation underdetermined.

Political arithmetic also joined observation with governing purposes. Which people were counted, which activities were valued, and which resources were compared reflected the questions of states and administrators. The resulting statistics should therefore be read as constructed instruments of inquiry. They can contain accurate observations while omitting conditions that were not important to their original producers.

Later demographic methods added more explicit sampling, life-table, and estimation procedures. The earlier achievement should not be described as a complete modern statistical system. Its historical significance is more specific: repeated records could be rearranged to expose regularities that were difficult to see in isolated cases. This expanded the scale of social evidence while creating a continuing need to inspect definitions, denominators, and coverage.

## **4. Moral statistics and the problem of normality**

André-Michel Guerry's *Essai sur la statistique morale de la France* appeared in 1833. Adolphe Quetelet's *Sur l'homme et le développement de ses facultés* followed in 1835. These works compared population regularities and helped make distributions of social phenomena objects of systematic inquiry. Guerry's tables and maps and Quetelet's discussion of statistical regularity provided ways to examine aggregate relationships beyond individual narratives. [6, 7]

Such work raised a question that remains central to diagnostics: what does a departure from a regular pattern mean? A rare observation can identify a recording error, a local condition, or an informative

exception. Statistical rarity by itself does not tell us which explanation is appropriate. Nor does the frequency of a condition establish that the condition is desirable.

The word normal is particularly difficult. It can refer to what is common, to the expectation of a model, to a prescribed standard, or to a judgment about acceptable conduct. These meanings need separate treatment. A society may regularly produce an unequal outcome. Its regularity does not turn that outcome into an ethical standard. Equally, an unfamiliar practice is not a defect merely because it differs from the analyst's reference population.

Recorded crime illustrates the problem. A criminal statistic depends on conduct, legal definitions, reporting, investigation, and classification. Two districts can differ because the underlying activities differ. They can also differ because one has more intensive enforcement or greater willingness to report. A map of recorded crime locates the result of this combined process. It does not automatically locate a stable property of the people living there.

An aggregate relationship introduces a further boundary. If districts with one characteristic also have a higher recorded rate, the relationship does not tell us which individuals account for the rate. Later work on ecological inference made this limitation explicit. At this stage of the history, the important point is that a new scale of observation created questions that individual case descriptions could not settle. Chapter 19 returns to the distinction between the individual and the aggregate.

We can therefore read moral statistics as both an expansion of comparative evidence and a warning about classifications. The historical tables are valuable artifacts. Their category names and social assumptions are also objects for examination. Our diagnostic reading preserves the numerical innovation without adopting the moral judgments that accompanied some of its classifications.

## **5. Social surveys, reform, and case diagnosis**

Harriet Martineau's *How to Observe: Morals and Manners*, published in 1838, treated social observation as an activity requiring preparation and disciplined interpretation. The observer needed to look beyond immediate impressions and examine relationships between institutions, everyday practices, and professed values. This is a useful starting point for a history in which observation itself becomes an object of criticism. [8]

Later urban surveys joined field observation with mapping and tabulation. Charles Booth's *London inquiry* ran from 1886 to 1903. Its poverty maps connected street-level classifications with notebooks and other records. The later maps are commonly identified with the 1898–1899 survey period, although the sheets were published in the survey volumes in 1902–1903. Keeping these dates separate helps us distinguish the observation from the publication. [9]

*Hull-House Maps and Papers*, published by the settlement's residents in 1895, connected maps of nationalities and wages with accounts of local working and living conditions. Du Bois's *The Philadelphia Negro*, published in 1899, likewise combined systematic investigation with attention to the historical

and institutional setting of a community. These works show why the development of empirical sociology cannot be represented only as a succession of abstract theories. Investigators constructed substantial bodies of evidence about particular places. [10, 11]

The scale of a survey creates its own diagnostic boundary. A street color summarizes a classification; it does not reproduce the circumstances of every household. A questionnaire makes cases comparable by asking similar questions, but it also selects which differences can be recorded. We need to move between the summary and the underlying record when a surprising result appears.

Richmond's Social Diagnosis made the assembly and assessment of social evidence explicit in 1917. Its casework setting required attention to a person's circumstances and relationships rather than the interpretation of one isolated statement. For our history, the work establishes a direct use of diagnosis in a social field. It should not be redescribed merely as an anticipation of software debugging. [1]

Survey and casework evidence also carries the position of the investigator. A household's explanation may be recorded within categories chosen by a reform organization, an employer, or an administrator. A useful diagnostic report therefore preserves the difference between the participant's statement and the analyst's classification. Otherwise, a later reader may mistake the institution's judgment for an unmediated observation.

These investigations connected description with proposed change. That connection increases the importance of checking the explanation. If poor outcomes are attributed to the wrong level, a reform may correct the reporting procedure while leaving the social condition intact, or alter people's circumstances when the apparent problem was created by the measurement. The evidence needs to support both the diagnosis and the level at which an intervention is proposed.

## PART II

## Explanation, observation, and the construction of measures

### 6. Social facts, meanings, and structural explanations

A description of hardship does not settle its explanation. Friedrich Engels's *The Condition of the Working Class in England*, first published in German in 1845, connected observations of industrial life with an account of economic and social organization. Its diagnostic relevance lies in shifting the explanatory level from an isolated person to the conditions within which people lived and worked. The work also has a political argument, which needs to be distinguished from each item of evidence used to support it. [12]

Marx's first volume of *Capital*, published in 1867, used material such as factory inspectors' reports in its analysis of the working day. Administrative evidence entered a criticism of the relations through which work and its products were organized. This illustrates a further use of social records: an institution's own documentation can be examined through a question that differs from the purpose for which it was produced. Acceptance or rejection of the larger theory should not substitute for examining the particular evidential connection. [13]

Durkheim's *The Rules of Sociological Method*, published in 1895, treated social facts as objects requiring systematic investigation. *Suicide*, published in 1897, developed comparative arguments from recorded rates. The move to social explanation made relationships between groups and institutions central to the inquiry. It also left a continuing problem: the categories in official records may be affected by the very institutions being compared. [14, 15]

Weber's 1904 essay on objectivity in social science addressed a different aspect of explanation. Researchers select problems through their interests and construct concepts for analysis. An ideal type accentuates selected relationships; it is not an average person or a complete copy of a society. The concept can support comparison precisely because its construction is explicit. Its usefulness still depends on the question and the evidence. [16]

These approaches should not be compressed into one method. A structural explanation asks how relationships and institutions generate conditions. A comparison of rates asks whether a proposed account is consistent with observed variation. Interpretive inquiry asks how action is oriented through meaning. A particular investigation may need more than one of these questions without making their answers interchangeable.

Consider a hypothetical organization with frequent staff departures. A table may show that departures concentrate in one department. Interviews may reveal how employees interpret a promotion rule.

Employment records may show that the rule systematically blocks a group's progression. The concentration, the interpretation, and the institutional mechanism are related evidence. None is simply a longer version of the others.

Our diagnostic task is to preserve the level of the claim. A group rate does not identify the motive of a particular individual. An individual's account does not establish the frequency of that motive in the population. A structural explanation needs evidence connecting the proposed relationship to the observed outcome. Otherwise, a general theory can become a label attached to any result after it occurs.

## **7. Ethnography and the preservation of context**

Bronislaw Malinowski's *Argonauts of the Western Pacific*, published in 1922, made fieldwork methods part of its account of exchange in the Trobriand region. Its introduction distinguished different kinds of material gathered through observation and participation. The investigator's presence, the detail of ordinary life, and participants' understandings all mattered to the resulting account. The colonial setting of this work also matters to how its authority and access were produced. [17]

An ethnographic record can preserve relationships that a predefined questionnaire does not anticipate. An apparently minor interruption may show who is permitted to speak. An exception to an official procedure may reveal how work actually gets done. Such observations become useful evidence when the field note records enough context for a later reader to understand the event and the observer's position.

Clifford Geertz's *The Interpretation of Cultures*, published in 1973, developed an interpretive approach associated with thick description. An observable movement acquires different significance according to the conventions and intentions within which it occurs. The diagnostic consequence is that recording the visible act alone may leave the event unidentified at the level relevant to the inquiry. [18]

There is a related distinction between generating an explanation and confirming it. Glaser and Strauss's *The Discovery of Grounded Theory*, published in 1967, described the development of theory through comparison with qualitative data. Categories can be revised as cases are examined. This gives the analyst a way to make conceptual change part of the research process rather than conceal it behind an apparently fixed scheme. [19]

A hypothetical field note illustrates the need for a preserved distinction. The note may say that a worker remained silent during a meeting. A later code may classify the event as agreement. Yet silence could also indicate uncertainty, exclusion, deference, fatigue, or a decision to object elsewhere. The code is an interpretation of the note. If the original record is discarded, the interpretation becomes difficult to challenge.

Ethnographic diagnosis therefore benefits from a history of interpretation. We need to know which observations prompted a category, which cases complicated it, and which alternatives were considered.

Participant disagreement may expose a mistaken interpretation, although agreement by one participant cannot settle every question about an institution. The analyst's account remains an argument supported by evidence.

The connection to trace analysis concerns context and reconstruction. A trace message can be read differently when its surrounding activity is known. Social events add the further problem that participants have their own interpretations and may contest the analyst's categories. This difference is part of the comparison, not an obstacle to be removed from it.

## **8. Probability sampling and the diagnosis of survey error**

Large surveys created a practical question: how can observations from a subset support statements about a wider population? Neyman's 1934 paper distinguished stratified sampling from purposive selection and developed an explicit probability basis for inference. The sample's evidential role depended on its selection procedure, not simply on its size or apparent resemblance to the population. [20]

Mahalanobis's 1944 paper on large-scale sample surveys examined design problems involving cost, precision, and the arrangement of sampling units. Its inclusion is important because the development of survey methodology was also a history of large empirical programs and methodological work in India. Sampling was a problem of organizing observation under material constraints as well as a problem of probability theory. [21]

Probability selection does not remove every source of error. A sampling frame may omit parts of the intended population. Selected people may not respond. A question may be misunderstood. An interviewer may change its wording. Responses may be coded incorrectly. Groves and Lyberg's 2010 account of total survey error brought this wider collection of problems into a historical and methodological review. [22]

We can think of the survey as a succession of transformations. A target population becomes a frame, a selected sample, a set of contacted people, a set of responses, and finally an analytical dataset. At each transition, cases may be lost or represented differently. The final table does not reveal these transformations unless the relevant records are retained.

Suppose an online questionnaire reports very high satisfaction with a public service. The first diagnostic question is whether the invitation reached people who could not complete the service online. If those users were excluded before selection, a larger number of completed questionnaires does not resolve the coverage problem. If they were invited but responded less often, the investigation concerns a different boundary.

Nonresponse rates are useful signals, but the amount of nonresponse alone does not determine the bias in every estimate. What matters is the relationship between response and the quantity being estimated, together with the adjustment information available. Weighting can address some imbalances under

explicit assumptions. It cannot create information about an omitted group when no suitable evidence exists.

The historical movement toward probability sampling therefore expanded both inference and diagnosis. Researchers could examine how an estimate depended on selection and could separate sampling variability from other defects. The distinction remains essential when a contemporary platform supplies millions of observations. A large trace collection still needs an account of who could enter it and how they became visible.

## 9. Psychological measurement and construct validation

Social scientists often investigate attributes that are not directly recorded as physical readings. Ability, trust, political efficacy, and perceived control become available through responses or actions interpreted as indicators. The diagnostic problem begins with that interpretation. An indicator can be measured precisely while its relationship to the intended construct remains uncertain.

The earlier history also includes an explicit connection between measurement and social hierarchy. Galton's *Inquiries into Human Faculty and Its Development*, published in 1883, joined investigations of human differences with a eugenic program. Its classifications and purposes are part of the history to be examined. A numerical measure cannot by itself justify ranking people's social worth or prescribing their opportunities. The methodological history must preserve this distinction instead of treating quantification as an automatic escape from prejudice. [23]

Spearman's 1904 work on general intelligence examined patterns of association between performances and helped establish the factor-analytic approach to underlying variation. Likert's 1932 monograph developed a technique for measuring attitudes through responses to sets of statements. These developments made relationships among observed responses central to the construction of measures. Neither a factor nor a summed score, however, carries its substantive interpretation automatically. [24, 25]

Cronbach and Meehl's 1955 article made construct validity a problem of the relationship between test interpretation, theory, and empirical evidence. A disappointing prediction could call into question the measure, an auxiliary assumption, or the theory in which the construct was used. Validation therefore required more than showing one favorable correlation. [26]

Campbell and Fiske's 1959 multitrait-multimethod matrix provided a way to examine whether observed agreement reflected the intended traits or the methods used to measure them. The distinction is diagnostic. Two questionnaires may agree partly because they share wording, response conventions, or a common source, rather than because they independently identify the same attribute. [27]

Bollen's 1989 *Structural Equations with Latent Variables* presented a developed framework connecting measurement models and structural relationships. Such models make assumptions explicit, including which indicators depend on which latent variables and how errors are related. A good fit to observed

associations does not by itself establish a unique causal explanation. Alternative specifications may reproduce similar patterns. [28]

Consider a hypothetical educational assessment used to diagnose difficulty with mathematical reasoning. A low score could reflect the intended skill, unfamiliar language, inaccessible presentation, or insufficient time. Repeating the same format can produce reliable scores while preserving the same limitation. A diagnostic evaluation needs evidence about the intended interpretation and use of the score, including the conditions under which that interpretation fails.

This history helps distinguish reliability from validity. Reliability concerns the consistency of a measurement under specified conditions. Validity concerns the support for an interpretation or use. Improving one does not guarantee the other. The distinction also applies to automated coding: a system can assign the same label consistently while systematically representing the wrong social phenomenon.

## **10. Comparability across groups, languages, and settings**

Comparison requires some continuity of meaning. If a survey item functions differently across groups, an observed difference in scores may combine a difference in the intended construct with a difference in the measurement process. Translation alone cannot establish that the same question occupies the same place in different social settings.

Meredith's 1993 treatment of measurement invariance and factorial invariance made conditions for group comparisons explicit within factor models. Different forms of invariance support different comparisons. Establishing a common pattern of relationships does not automatically justify comparing latent means. The analyst must identify the comparison required and examine the assumptions relevant to it. [29]

The population from which a result is obtained also matters. Henrich, Heine, and Norenzayan's 2010 discussion of WEIRD samples challenged broad claims about human behavior based on narrow populations. Their argument concerns the empirical reach of generalizations. It does not mean that a result obtained in a Western, educated, industrialized, rich, and democratic setting is intrinsically false. [30]

A hypothetical question about trust in neighbors illustrates both problems. In one setting, neighbor may imply daily reciprocal obligations. In another, it may refer mainly to residential proximity. A numerical difference can be real at the level of the answers while remaining ambiguous at the level of the intended construct. Interviews, translation review, and alternative indicators may help locate the boundary.

Comparability across time presents similar difficulties. A qualification can change its labor-market meaning. A household category can be redefined. An institution may move a service online. A stable variable name in a dataset does not establish a stable measurement process. We need a version history of questions, coding rules, and administrative practices.

This gives us another interpretation of a diagnostic invariant. We seek a relationship that remains meaningful under the comparisons we intend to make. The invariant is not assumed because a column has kept its name. It is supported by evidence about the production and interpretation of its values. When that support fails, the appropriate repair may be a revised comparison rather than a corrected number.

Cross-cultural work consequently expands the evidence required for diagnosis. The issue is not only how much data was collected, but whether the same inferential step is warranted in each setting. A method can be transported as a procedure while the substantive meaning of its output changes. A useful report records both the common procedure and the limits of the comparison.

## PART III

## Experiments, causal inference, and historical comparison

### 11. Experimental design and rival explanations

An experiment changes the conditions under which evidence is produced. Instead of observing only the arrangements that happen to exist, an investigator assigns a treatment or organizes a comparison. This can make some rival explanations less plausible. It also introduces new questions about compliance, measurement, and the setting created by the experiment.

Fisher's *The Design of Experiments*, first published in 1935, systematized principles including randomization and replication. Its agricultural setting is a reminder that methods used in social sciences often developed across disciplinary boundaries. Random assignment provides a basis for evaluating treatment comparisons through the assignment mechanism. It does not guarantee that every realized group is identical on every characteristic. [31]

Campbell and Stanley's 1963 chapter on experimental and quasi-experimental designs organized threats to the interpretation of observed changes. History, maturation, selection, testing, and changes in measurement could produce patterns mistaken for treatment effects. The taxonomy is diagnostic because it connects an observed result with alternative processes that may have produced it. [32]

Consider a hypothetical school that introduces a reading program after an unusually poor assessment. Scores rise at the next assessment. The change may include an effect of the program, but it may also include regression toward more typical performance, changes in the pupils tested, or increased familiarity with the assessment. The before-and-after difference is an observation requiring an explanation.

Randomized assignment can strengthen that explanation when the assignment is implemented and analyzed appropriately. However, the recorded treatment may differ from what participants receive. Teachers can adapt the program, pupils can move between groups, and information can pass from treated to untreated classes. An investigation of the assignment, implementation, and outcome may therefore require different artifacts.

The experimental setting also affects interpretation. Participants may respond to being observed, to expectations about the study, or to the presence of additional staff. We should not invoke a universal observation effect whenever a result is inconvenient. We need evidence about the particular relationship between the study procedures and the behavior being measured.

A well-designed experiment narrows a causal question. It does not settle every question about meaning, fairness, institutional feasibility, or transfer to another population. The result may show an average effect under the conditions studied while leaving the mechanism and distribution of benefits uncertain.

These are additional investigations rather than reasons to discard the experiment's more specific contribution.

## 12. Econometrics and the identification of a model

Economic records contain many variables that change together. Prices, employment, income, and production are jointly shaped by decisions and institutions. A fitted relationship can describe these movements without identifying the effect of an intervention. The historical development of econometrics made this distinction increasingly explicit.

Haavelmo's 1944 *The Probability Approach in Econometrics* connected economic models with probabilistic assumptions about observable data. It treated the relationship between theory, measurement, and statistical inference as a central methodological problem. The work helped establish a framework in which an economic model could be examined through the distribution of the evidence it implied. [33]

A market example shows why this matters. Observed prices and quantities can reflect movements in supply, demand, or both. A line fitted to their joint observations does not automatically recover one underlying behavioral relationship. Additional restrictions or sources of variation are needed to distinguish explanations. This is an identification problem, not merely a problem of calculating the fitted line more accurately.

We should also distinguish identification from estimation. Identification asks whether the assumptions and available observations determine the target quantity. Estimation concerns learning that quantity from a finite dataset. More observations can improve estimation while leaving an identification problem unresolved. An algorithm can converge precisely to a result that does not answer the intended causal question.

Residual analysis adds another diagnostic level. Patterns left after fitting a model may indicate an omitted relationship, a wrong functional form, an influential observation, or a change in the data-generating process. The pattern signals that the model requires examination. It does not identify a unique repair. Chapter 28 develops this point in connection with missing data and sensitivity analysis.

In social research, the model's categories also deserve inspection. A variable called income may refer to earnings, disposable household income, or reported income in a particular administrative system. Combining these without a documented transformation can create an apparently technical disagreement that originates in a difference of meaning.

Econometric diagnosis therefore has several connected objects: the economic account, the measurement procedure, the statistical model, and its computational implementation. Correcting the implementation is valuable, but it does not establish the other relationships. The history of increasing computational power is accompanied by a continuing need to state what the model identifies and which assumptions carry that identification.

### 13. Potential outcomes, causal graphs, and observational comparisons

The effect of an intervention concerns a comparison that cannot usually be observed for the same unit under both alternatives at the same time. Rubin's 1974 article on randomized and nonrandomized studies formulated treatment effects through potential outcomes and considered the conditions under which different study designs could estimate them. This made the missing comparison an explicit part of the problem. [34]

Suppose we want to know whether a training program improves employment. The outcome for a participant after training is observable. The outcome that the same person would have experienced without training, under the specified alternative conditions, is not observed. Comparing participants with nonparticipants therefore requires an account of how they differ before participation and how those differences affect outcomes.

Rosenbaum and Rubin's 1983 propensity-score work showed how treatment assignment probabilities could support adjustment for observed covariates. The causal interpretation requires assumptions, including appropriate overlap and the absence of unmeasured confounding conditional on the covariates used. Balancing recorded characteristics does not demonstrate that unrecorded differences have disappeared. [35]

Pearl's Causality, first published in 2000 and expanded in a second edition in 2009, developed graphical and structural approaches to causal questions. A causal graph makes proposed relationships explicit and helps determine which adjustments are appropriate under those relationships. The graph is a representation of causal assumptions. Observed correlations alone do not generally determine its correct form. [36]

The diagnostic value of a graph includes exposing a mistaken adjustment. A variable may be a consequence of treatment, a common cause, or a common effect of other variables. Treating all available variables as interchangeable controls can change the question or introduce bias. The analyst needs a causal account of the variable's position before using it to repair an apparent imbalance.

These approaches clarify the difference between predicting an outcome and estimating an intervention effect. A feature can improve prediction while being unsuitable for a particular causal adjustment. A model can predict employment well because it uses information generated after training, yet provide no valid estimate of the training effect.

For diagnostic reporting, the useful artifact is consequently more than a coefficient table. It includes the intervention being compared, the population, the outcome definition, the assumed causal structure or assignment conditions, and the checks made possible by the data. Some assumptions remain only partly testable. They should remain visible as assumptions rather than being converted into facts by the appearance of a precise estimate.

## 14. Quasi-experiments and naturally occurring comparisons

Many social interventions cannot be randomized by the investigator. Researchers therefore examine variation produced by rules, thresholds, timing, or external events. Such variation can support a causal comparison when its relationship to the outcome is understood. Calling an event natural does not itself establish the required independence.

Thistlethwaite and Campbell's 1960 regression-discontinuity article examined a design in which treatment assignment changes at a threshold. Observations close to the boundary can support a local comparison under conditions about continuity and assignment. A threshold is useful because it creates a particular structure in the evidence. It is not a general guarantee that people on opposite sides are interchangeable. [37]

Card and Krueger's 1994 study of fast-food employment compared changes in New Jersey and Pennsylvania around New Jersey's minimum-wage increase. It became a prominent example of using a comparison group and changes over time. The historical case should retain its particular setting and research design; it does not supply a universal conclusion about every minimum-wage change. [38]

In a difference-in-differences analysis, the counterfactual comparison concerns how the treated group would have changed without treatment. Pre-intervention trends can provide useful evidence, but similar earlier trends do not prove that the counterfactual later trends would have remained parallel. Anticipation, concurrent policies, and changes in group composition can affect the comparison.

Instrumental-variable methods address another structure. Angrist, Imbens, and Rubin's 1996 analysis showed how, under assumptions such as exclusion and monotonicity, an instrumental-variable estimand can identify an average causal effect for a particular subgroup of compliers. The target is therefore not automatically the effect for every person in the population. [39]

These designs have characteristic diagnostic questions. Can people manipulate the assignment threshold? Did a policy affect the comparison group? Does the proposed instrument change the outcome through another route? Did the definition of the outcome change when treatment began? Each question points toward different evidence rather than another generic robustness check.

The history of quasi-experimental methods is a history of making a comparison's justification more explicit. A good investigation explains why the available variation is informative and what it cannot establish. When an assumption is doubtful, the appropriate result may be a restricted claim or a sensitivity analysis. Precision in a statistical calculation does not remove uncertainty about the comparison that gives it causal meaning.

## 15. Comparative cases and process tracing

Historical and political inquiries often concern a small number of complex cases. A revolution, a policy failure, or the transformation of an institution cannot always be represented adequately as one row in a

large table. Researchers need methods for examining configurations and sequences while retaining the evidence that supports the explanation.

Mill's *A System of Logic*, first published in 1843, set out methods of agreement, difference, and related forms of inquiry. These methods remain useful for understanding the logic of comparison. Their application to social cases requires care because cases usually differ in many respects, causes can operate in combinations, and relevant conditions may be omitted from the comparison. [40]

Ragin's *The Comparative Method*, published in 1987, developed a set-theoretic approach to configurations of conditions. Different combinations can be associated with the same outcome. This helps preserve forms of causal complexity that a search for one isolated factor may overlook. The result still depends on case selection, the calibration of categories, and the treatment of configurations absent from the observed evidence. [41]

Process tracing examines evidence within a case. Collier's 2011 account emphasized careful description and diagnostic evidence in relation to hypotheses. A sequence becomes explanatory when the evidence bears on the proposed process connecting conditions and outcomes. Listing events in date order does not by itself establish such a process. [42]

Consider a hypothetical account of a policy reversal. A public announcement, an internal memorandum, and a later recollection may suggest different decision sequences. The analyst needs to establish when each artifact was produced, what its author could know, and whether it records a decision or a justification prepared afterward. The provenance of the message can matter as much as its explicit content.

Terms such as hoop test and smoking-gun test describe different relationships between evidence and hypotheses. Their force depends on the assumptions that make an observation necessary, expected, or unusual under a rival explanation. A document is not decisive simply because it is vivid or confidential. Its diagnostic value follows from how it changes the relative support for the hypotheses being examined.

The connection to trace analysis is close but bounded. Both activities can reconstruct an event sequence from partial artifacts. Social process tracing must additionally examine strategic accounts, changing meanings, and evidence deliberately produced for an audience. A credible reconstruction retains these conditions and shows where the sequence is observed, inferred, or still uncertain.

## PART IV

## Indicators, sequences, networks, and organizations

### 16. National accounts and social indicators

An institution cannot respond to every individual record separately. It uses indicators to summarize activity and compare periods or populations. This creates a diagnostic instrument, but it also creates a boundary between the condition of interest and the quantity selected to represent it.

Kuznets's National Income, 1929–1932, published in 1934, formed part of the development of systematic national economic measurement. National accounts made relationships between economic aggregates available for repeated comparison. The definition of an aggregate determines which activities enter it and how they are valued. An increase in measured production does not, by itself, describe the distribution or quality of people's lives. [43]

The first Human Development Report, published by UNDP in 1990, organized development around an expansion of human opportunities and introduced the Human Development Index. Sen's Development as Freedom, published in 1999, developed an account centered on people's substantive freedoms and capabilities. These contributions broadened the question being evaluated beyond income alone. They did not remove the choices involved in selecting and combining indicators. [44, 45]

A composite index can make several dimensions visible in one summary. It can also conceal the route to a change. Two regions may receive the same score through different combinations of health, education, and income. A diagnostic comparison therefore returns to the components when the summary matters for explanation or action.

Revisions are another source of apparent discontinuity. An agency may improve coverage, revise estimates, change a base year, or redefine a category. A published time series can consequently differ from the series available to decision-makers at an earlier date. We need to distinguish the history of the phenomenon from the history of its estimates.

Consider a hypothetical dashboard of social inclusion. The dashboard reports improvement because more people have registered for a service. Registration may be a useful measure of reach. It does not establish that people can use the service effectively or that their circumstances have improved. The indicator's name should not carry a stronger claim than its production procedure supports.

The historical expansion of social indicators made new forms of comparison possible. It also made indicator design a substantive part of social inquiry. A diagnostic report should specify the construct, the observable components, their combination, and the purpose of the comparison. This allows a reader to decide whether a surprising value is a change in social conditions, a change in measurement, or a mismatch between the measure and the question.

## 17. Longitudinal studies and histories of transition

A cross-sectional observation provides a view of a population at one time. It does not show how particular people reached their current positions. Longitudinal research preserves repeated observations and makes transitions available for analysis. The evidence can then distinguish persistent conditions from movement into and out of them.

The Panel Study of Income Dynamics began in the United States in 1968. Its continuing household observations created a resource for examining income, family, and related changes over time. Its historical importance lies in the preservation of linked observations, not simply in the accumulation of more survey rows. The identity and relationships of the units connect one observation with another. [46]

Event-history methods ask about the timing of a transition and the conditions under which it occurs. Cox's 1972 paper on regression models and life-tables provided a widely used model for event rates with covariates. Its application extends across disciplinary boundaries. In a social investigation, the analyst must still define the event, the period at risk, and the meaning of observations that end before the event is observed. [47]

Abbott's 1995 review of sequence analysis distinguished approaches to steps and transitions from approaches to whole sequences. The order of events can carry information that a collection of durations does not preserve. Two people may spend equal time in education, employment, and unemployment while following different trajectories. [48]

Missingness has a temporal structure as well. A participant may leave a panel because of migration, illness, administrative difficulty, or unwillingness to continue. If these processes relate to the outcome, treating the remaining observations as an unchanged population can distort the history. A gap should retain a status that distinguishes an unobserved interval from a recorded absence of the event.

Retrospective accounts create different uncertainties. People may remember an event but place it in the wrong month or connect it with an explanation formed later. Administrative timestamps may record data entry rather than the time of the underlying activity. Ordering two records by their stored time can therefore produce an incorrect sequence.

These problems are familiar from trace analysis, where message time, event time, and collection time can differ. In social research, the reconstruction also concerns memory, changing categories, and the continuity of persons and households. A useful longitudinal artifact retains identifiers, transitions, interval definitions, and revisions. The additional structure makes it possible to ask whether a discontinuity belongs to the life history or to its recording.

## 18. Sociometry and social network evidence

Many social explanations concern relationships rather than isolated attributes. A person's access to information, support, or influence can depend on whom they know and how those people are connected. Network representations make selected relationships available for inspection and calculation.

Moreno's *Who Shall Survive?*, published in 1934, developed sociometric approaches to human interrelations. The sociogram offered a way to represent choices and connections within groups. The representation brought structural features into view, including separation, concentration, and relations between subgroups. Its meaning depended on the question used to elicit a relationship. A choice of preferred collaborator is not identical to a record of actual collaboration. [49]

Granovetter's 1973 *The Strength of Weak Ties* connected interpersonal relationships with wider social structure. A tie's role can depend on how it links otherwise separated parts of a network. This shifted attention from the amount of contact alone to the arrangement of contact. [50]

A network is therefore a model of a particular relationship. A friendship network, a money-transfer network, and a network of mentions in public messages are different artifacts even when their nodes represent the same people. Combining them without preserving edge meaning can create a diagram that is visually persuasive and analytically ambiguous.

Network boundaries also affect diagnosis. An apparently isolated person may have important ties outside the observed institution. A central actor may appear central because their communications are recorded more completely. Missing edges can change the apparent route through which information could have traveled. The absence of a recorded tie is not always evidence that no relationship exists.

Observed similarity between connected people presents another problem. It can result from influence, selection of similar contacts, or shared conditions. The network makes the association visible, but a causal account requires additional evidence. The same distinction applies when an online diffusion diagram is described as a mechanism merely because its arrows resemble a plausible story.

Consider a hypothetical service network in which several agencies exchange referrals. A concentration of delayed cases at one agency may reflect its workload, the complexity of the cases it receives, or a bottleneck in a shared external service. Network structure helps locate useful comparisons. It does not determine the explanation without evidence about the activities carried by the links.

## 19. Spatial evidence, aggregation, and ecological inference

A map makes the location of recorded differences visible. It can reveal concentrations that are difficult to recognize in a table. However, the visual unit is often an area, while the explanation concerns people or events within it. This difference creates a diagnostic boundary.

Robinson's 1950 article on ecological correlations demonstrated that relationships between aggregate quantities can differ from relationships between individual characteristics. An association across regions cannot simply be transferred to individuals living in those regions. The problem concerns the level of inference, even when every aggregate number is calculated correctly. [51]

Changing the boundaries of areas can also change the observed pattern. Suppose a city combines a high-income and a low-income neighborhood into one reporting district. The district average can conceal both local conditions. A later reorganization may produce an apparent increase in inequality between districts without any household changing its income. The map's geography is part of the measurement procedure.

Anselin's 1995 work on Local Indicators of Spatial Association provided methods for examining local spatial patterns. Such statistics can identify locations that contribute to clustering or appear unusual relative to neighboring observations. Interpretation depends on the specified neighborhood relationships and on the inferential procedure. A local cluster is a prompt for explanation rather than a complete account of its origin. [52]

Spatial records also reflect institutional presence. Recorded incidents may concentrate where reporting facilities are accessible or where enforcement is intensive. A map of service use may resemble a map of service availability. The analyst needs to ask whether the spatial pattern belongs to the underlying condition, the observation process, or both.

Criminology makes this particularly visible. A reduction in recorded offenses within a treated area can coexist with displacement to another area, a change in reporting, or a change in police classification. Evidence from neighboring locations and other sources can help examine these alternatives. It should not be assumed that a favorable local indicator describes the entire relevant system.

Spatial diagnosis consequently benefits from multiple scales and representations. We may compare points, areas, routes, and networks, provided the unit of inference remains explicit. The comparison resembles changing a viewpoint in trace analysis. Each view exposes selected relationships, while the transformation between views needs to preserve enough information for the explanation being tested.

## **20. Interaction records and organizational failure**

Organizations produce official descriptions of their procedures. Actual work also depends on conversations, negotiated exceptions, informal coordination, and practical interpretations. The relationship between the two is a recurring object of social investigation. A process diagram can be accurate as a rule while incomplete as an account of what happens.

Garfinkel's *Studies in Ethnomethodology*, published in 1967, examined how people produce and recognize an intelligible social order in ordinary activity. The analyst's categories are themselves part of this problem. A record described as an objective account may rely on practical judgments that participants make in creating and using it. [53]

Sacks, Schegloff, and Jefferson's 1974 paper on turn-taking in conversation developed a model grounded in observable features of conversational organization. Recordings and detailed transcripts made sequences available for repeated inspection. Pauses, overlaps, and the placement of a response can matter to the interpretation of an interaction. A summary of what was said may lose this structure. [54]

Organizational diagnosis can use such evidence to examine handoffs. Suppose one office records a request as transferred while the receiving office records it as awaiting clarification. Both entries may be locally correct. The failure concerns the relationship between the offices' definitions of completion. The useful artifact is the sequence connecting the records, including the conversation through which responsibility was negotiated.

Perrow's Normal Accidents, first published in 1984, examined technological accidents through organizational relationships, especially complex interactions and tight coupling. Its contribution connects this history with technical and hardware diagnostics. Failures can emerge from the arrangement of activities and dependencies even when an account centered on one defective component is inadequate. [55]

This does not mean that every accident is inevitable or that component analysis is unimportant. It means that the explanatory level should follow the evidence. A local error may be enabled, amplified, or left undetected by the surrounding organization. Correcting the last visible action can leave those relationships intact.

An organizational postmortem therefore needs more than a final incident narrative. It needs records of what participants could see, which responsibilities they understood, and how conditions changed over time. The distinction between knowledge available during the event and knowledge available afterward is essential. Otherwise, the reconstructed history can make uncertain decisions appear obviously mistaken and conceal the diagnostic limits under which people acted.

**PART V****Intervention, institutional diagnosis, and reflexivity****21. Program evaluation and the explanation of outcomes**

An evaluation asks whether a program produced an outcome and what that outcome means for its purpose. This differs from counting activities. A program may complete every scheduled session while failing to change the condition it was intended to address. It may also produce a useful effect through a route that differs from its original explanation.

The experimental and quasi-experimental tradition contributed methods for estimating effects, while theory-oriented evaluation examined the relationships expected to connect activities and outcomes. Pawson and Tilley's *Realistic Evaluation*, published in 1997, placed mechanisms and their contexts at the center of explanation. The question concerns how a program's resources and participants' responses produce outcomes under particular conditions. [56]

Consider a hypothetical employment service. Its initial explanation is that interview coaching improves employment. Attendance records show that the sessions occurred, and employment later increases. Interviews indicate that the service also connected participants with employers through personal introductions. The outcome may be favorable, but the original mechanism has not been established merely because the program and the improvement coexist.

Implementation evidence is therefore distinct from outcome evidence. A manual describes the intended procedure. A delivery record shows selected activities. Participant accounts reveal how the activities were experienced. An outcome measure records a consequence under its own definition. Disagreement between these artifacts can expose a useful diagnostic boundary.

Variation between sites should also be examined before it is treated as noise. Different staffing, eligibility rules, local labor markets, or participant needs can change how a program operates. A common intervention label may conceal several effective treatments. Conversely, locally different practices may realize the same underlying mechanism.

Evaluation also needs an explicit time horizon. A short-term increase in activity may precede a later benefit or a later cost. An effect can disappear when external support is withdrawn. The point at which measurement stops is therefore part of the account of success. It is not merely an administrative detail of the study.

The diagnostic connection is the separation of failure modes. A program can fail because its explanation is wrong, its implementation differs from the design, its context blocks the proposed mechanism, or its evaluation does not measure the intended outcome. These possibilities are related, but they do not call for the same repair. An effective evaluation makes the distinctions available before recommending continuation, modification, or replacement.

## 22. Action research and participation in the investigation

In action research, inquiry is connected with attempts to change a situation. Lewin's 1946 Action Research and Minority Problems described research in relation to planning, action, and examination of results. The activity creates successive opportunities to revise both the intervention and the understanding of the problem. [57]

Participation changes who can contribute to that understanding. Freire's Pedagogy of the Oppressed, published in English in 1970, treated dialogue and critical reflection as central to education and action. It questioned an arrangement in which knowledge is simply deposited in passive recipients. For diagnostic inquiry, this raises a question about whose problem description organizes the work. [58]

Participatory action research also developed through experiences in Latin America, South Asia, and other settings. Fals-Borda and Rahman's 1991 collection Action and Knowledge examined the relationship between inquiry, action, and control over knowledge. Its contribution is not captured by describing participants only as additional sources of data. Participation can include influence over the questions, interpretation, and use of the results. [59]

Suppose an investigator defines a low uptake of a service as a failure to communicate its availability. Participants may describe a different problem: the service's eligibility checks impose costs that exceed its benefit. A communication intervention could then increase awareness without addressing the reason for nonparticipation. Including participants in the initial diagnosis can change the object of the investigation.

Participation is not a guarantee of agreement or accuracy. Communities contain different positions, and the people most able to attend meetings may not represent those most affected. A local consensus can exclude dissent. The investigator still needs a record of who participated, what alternatives were considered, and how disagreements influenced the final account.

Repeated intervention also changes the evidence. After the first cycle, participants may have new expectations, relationships, and knowledge of the process. Later observations are not repetitions under unchanged conditions. We need to preserve the history of the intervention as part of the explanation of what follows.

The connection to debugging is consequently limited but useful. We can examine a proposed change, observe its consequences, and revise the account of the problem. We cannot assume that the desired social outcome is already specified or that the investigator alone has authority to define it. In participatory inquiry, the specification itself may become an object of collective examination.

## 23. Institutions and diagnosis beyond universal remedies

An institution organizes expectations, permissions, responsibilities, and sanctions. Its effects depend on how these arrangements interact with the activities they govern. A rule that works in one setting may fail in another because the relevant resources, relationships, or opportunities differ.

Ostrom's *Governing the Commons*, published in 1990, examined arrangements through which people managed common-pool resources. Its design principles drew attention to features of enduring institutions, including boundaries, monitoring, and procedures for dealing with violations and conflicts. They were part of a comparative analysis, not a mechanical recipe for every community. [60]

Ostrom's 2007 diagnostic approach made the rejection of universal remedies explicit. It proposed examining linked social and ecological systems through a framework of related variables at different levels. The purpose was to support more discriminating inquiry into why particular arrangements and outcomes occur. A framework organizes questions; it does not supply a causal answer before the case is studied. [2]

Scott's *Seeing Like a State*, published in 1998, examined the problems that can follow when administrative simplification supports ambitious schemes while disregarding local knowledge and practice. For our purposes, the relevant diagnostic boundary is between an arrangement as represented by an authority and the relationships through which it operates. [61]

Consider a hypothetical irrigation scheme. A central record assigns equal water entitlements to all users. The actual delivery depends on location, maintenance, timing, and the ability to challenge a violation. Formal equality in the register is not enough to establish equal access. Diagnosis requires the rules and evidence of their operation.

Monitoring is itself an institutional arrangement. Who observes, who receives the report, and what happens after a discrepancy is found affect the value of the evidence. A complete record that nobody can act upon may provide little practical correction. A powerful monitoring system without a means to contest mistaken records may create another source of harm.

This chapter provides one of the strongest historical connections with pattern-oriented diagnosis. Recurring arrangements can direct attention toward relevant relationships without determining the answer. A pattern of failure suggests questions about boundaries, incentives, communication, or enforcement. The local evidence then determines whether the pattern is applicable and what change is justified.

## 24. Reflexivity, predictions, and performance targets

Social descriptions can enter the processes they describe. People may act on a forecast, prepare for an assessment, or change their behavior to satisfy an indicator. The relationship between observation and action is then part of the system under investigation.

Merton's 1948 article on the self-fulfilling prophecy examined how a definition of a situation can help produce conduct that makes the originally false definition appear true. This gives a specific mechanism through which belief and action interact. It should not be generalized into the claim that every prediction creates its own outcome. [62]

Lucas's 1976 critique of econometric policy evaluation addressed another form of response. Relationships estimated under one policy arrangement may change when policy changes because people adapt their decisions. The issue concerns the stability of the behavioral relationships used for policy analysis. A historically successful predictive equation may not remain suitable for evaluating a new regime. [63]

Campbell's discussion of planned social change, published as a journal article in 1979, warned about the pressure placed on quantitative indicators when they become important for decision-making. Targets can change both conduct and recording. The resulting indicator may improve while its relationship to the original objective weakens. [64]

A hypothetical school can increase the reported proportion of successful pupils by improving teaching. It can also change entry rules, concentrate on the measured tasks, or alter which pupils are included in the reported denominator. These processes can coexist. An investigation that begins with an accusation of manipulation may miss real improvement; one that accepts the indicator without examining its production may miss the changed boundary.

Reflexivity also affects forecasts. A warning can prompt action that prevents the forecast outcome. The absence of the outcome does not necessarily show that the original warning was unreasonable. We need to examine the intervention induced by the warning and the counterfactual claim that was actually made.

The diagnostic artifact therefore needs to include the use of the measure. A survey answered privately for research is different from a form linked to a benefit decision, even when the questions are identical. The incentive structure can change the meaning of a response. Observability in social settings depends partly on what people expect observation to do.

## **25. Authority, ethics, and the ownership of evidence**

The right to observe and classify is not distributed equally. A researcher can have access to an institution's records while the people described in them have little ability to correct a mistake. A history of social evidence therefore also concerns authority over the questions, categories, and consequences of inquiry.

The Belmont Report, completed in 1978 and published in the Federal Register in 1979, articulated respect for persons, beneficence, and justice in research involving human participants. In this history, it represents an explicit effort to connect research procedures with obligations toward the people who

make the evidence possible. It is a historical ethical framework, not a complete account of every contemporary jurisdiction's requirements. [65]

Linda Tuhiwai Smith's *Decolonizing Methodologies*, first published in 1999, examined research through its relationship with colonial power and Indigenous peoples. The question extends beyond whether a particular observation is accurate. It concerns who defines the problem, whose knowledge is recognized, and who benefits from the resulting account. [66]

The CARE Principles for Indigenous Data Governance, released in 2019 and described by Carroll and colleagues in 2020, emphasize collective benefit, authority to control, responsibility, and ethics. D'Ignazio and Klein's *Data Feminism*, also published in 2020, connects data practices with the examination of power, context, and exclusion. These contributions make governance and perspective part of the analysis of data rather than an afterthought to collection. [67, 68]

A hypothetical dataset of community service use may contain correct entries while enabling an interpretation that stigmatizes the community. Removing names does not necessarily remove every collective consequence. Conversely, restricting access can make it harder to challenge an institution's account. These tensions require attention to the purpose and structure of access, not a universal rule that more visibility is always better.

Evidence can also be corrected through contestation. A participant may identify a mistranslation, a misleading category, or an event missing from an administrative history. A correction procedure improves the artifact's diagnostic value when the original record and the reason for revision remain distinguishable. Silent overwriting can erase the evidence needed to understand how the mistake occurred.

The useful connection with diagnostics is therefore an accountable relationship between record and claim. We should be able to ask who supplied the evidence, who transformed it, who can challenge it, and what decisions depend on it. Technical accuracy matters within that relationship. It cannot determine by itself whether the investigation's purpose or use is justified.

## PART VI

## Computational evidence and the examination of research

### 26. Digital traces and computational social science

Digital systems record transactions, searches, movements, and interactions as part of their operation. These records can make aspects of social activity visible at a scale and temporal resolution difficult to obtain through interviews alone. They also introduce the observation policies and business arrangements of the systems that produce them.

Lazer and colleagues' 2009 article on computational social science described research opportunities created by extensive digital records and computational methods. It also raised questions about access and the organization of research. The historical development concerned a new relationship between social inquiry and data-producing infrastructures, rather than simply a faster way to analyze an ordinary survey. [69]

Blumenstock, Cadamuro, and On's 2015 study combined mobile-phone metadata with survey information to predict economic characteristics. It provides a concrete example of linking a digital behavioral trace to a socially meaningful measure. The predictive relationship is learned through the reference evidence and the population studied. It should not be treated as a direct reading of wealth from a phone record. [70]

The provenance of a digital trace is therefore essential. A recorded click may reflect a person's intention, the interface's arrangement, an automated process, or several of these together. A platform's ranking system influences what can be seen and hence what can be clicked. Observed behavior is partly produced within the instrument through which it is observed.

The 2014 account of Google Flu by Lazer and colleagues examined the limitations of a prominent search-based prediction system. Although the target was epidemiological, the case is relevant here because the evidence consisted of human behavior mediated by a changing digital platform. A relationship between searches and an external condition can change when behavior, publicity, or the platform changes. [71]

Access creates another boundary. A researcher may receive an extract without knowing every filtering rule applied upstream. Deleted content, private activity, API limits, and changed definitions can alter what enters the dataset. A stable analytical script cannot compensate for an undocumented change in the source.

Computational social science consequently inherits the older problems of coverage, measurement, and interpretation. It adds detailed records of interaction with software systems. The diagnostic task is to connect the social event, the platform event, the stored record, and the analytical feature. When these are treated as one object, a technically accurate analysis can answer a different question from the one its title suggests.

## 27. Simulation and the distinction between generation and explanation

Simulation constructs a system whose behavior can be examined under specified rules. It is particularly useful when interactions produce aggregate outcomes that are difficult to anticipate from individual rules alone. The resulting trace is evidence about the constructed model. A further argument is required to connect that model with a social setting.

Schelling's 1971 *Dynamic Models of Segregation* showed how local preferences and movement could generate spatial separation. The models made a possible relationship between individual decisions and aggregate structure explicit. They did not demonstrate that one simplified mechanism explains every historical instance of residential segregation. Housing institutions, discrimination, economic constraints, and other processes require their own evidence. [72]

Epstein and Axtell's *Growing Artificial Societies*, published in 1996, developed the Sugarscape model as an approach to generating social patterns from interacting agents. The work made construction and repeated execution central to the inquiry. It helped show how a model could produce a history rather than only a static equilibrium. [73]

Simulation introduces a literal debugging problem. The implemented rules may differ from the written model. An update order may privilege some agents. A boundary condition may produce an unintended concentration. A random-number seed can affect one realization. The investigator needs checks that distinguish a programming defect from behavior implied by the intended rules.

Verification and validation answer different questions. Verification examines whether the implementation realizes the specified model. Validation examines whether the model is adequate for an intended empirical use. A verified simulation can be a poor representation of the social process. A visually convincing output can be produced by several different models.

Sensitivity analysis helps examine how results depend on parameters and structural choices. The investigation should include choices such as who can interact, what information agents possess, and whether learning or institutions change. These assumptions may matter more than small adjustments to one numerical parameter.

Consider a hypothetical simulation of service queues. The model reproduces an observed waiting-time distribution, but one version assumes limited staff and another assumes repeated rework. The same aggregate pattern does not identify which explanation applies to the institution. Records of staffing and case transitions are needed to distinguish them.

Simulation therefore creates a controlled setting for exploring mechanisms and their consequences. Its strongest diagnostic role is often to make an explanation precise enough to generate expectations that can be compared with evidence. The generated trace extends the argument; it does not replace observations of the world to which the argument is applied.

## 28. Residuals, missing data, and sensitivity to analytical choices

Statistical models produce diagnostic artifacts of their own. Residuals, influence measures, convergence reports, and alternative estimates make aspects of an analysis inspectable. Their interpretation requires attention to what each artifact tests and which assumptions remain outside it.

Cook's 1977 paper on influential observations provided a way to examine how individual observations affect a fitted regression. An influential case may be an error, but it may also be a valid and substantively important part of the population. Removing it solely because the conclusion changes would confuse the detection of dependence with a justification for exclusion. [74]

Rubin's 1976 *Inference and Missing Data* formalized relationships between the missingness mechanism and the analysis. The paper distinguished data missing at random from data observed at random and specified conditions under which the missingness process could be ignored. In particular, missing at random does not mean that missingness is unrelated to everything. It allows dependence on observed information within the specified framework. [75]

A hypothetical survey of income illustrates the issue. If nonresponse depends on income even after the available adjustment variables are considered, imputing from the observed cases under an inappropriate assumption can distort the result. Several completed datasets may quantify uncertainty within the imputation model while leaving uncertainty about that model's assumptions unresolved.

Sensitivity analysis makes some of these dependencies visible. We can ask how conclusions change under alternative plausible coding rules, missingness assumptions, specifications, or sample definitions. The purpose is to examine the support for an inference, not to search until one preferred result becomes statistically favorable.

Silberzahn and colleagues' 2018 *Many Analysts, One Data Set* study gave 29 teams the same data and research question concerning soccer referees' red-card decisions and players' skin tone. Variation in analytical choices produced variation in results. The study made the analysis procedure itself an empirical object. It did not imply that every possible analysis is equally defensible. [76]

The diagnostic response is a history of transformations and reasons. A reader should be able to see why cases were excluded, how variables were constructed, and which alternatives were examined. A clean final table without this history may hide the point at which reasonable analysts diverged.

Model diagnostics thus operate at several levels. A residual pattern concerns agreement with a fitted model. A sensitivity analysis concerns dependence on choices or assumptions. A substantive challenge concerns whether the model represents the intended problem. These artifacts can inform one another, but no single successful check establishes the correctness of the entire investigation.

## 29. Replication, preregistration, and reproducible analysis

Social research also examines its own capacity to produce dependable knowledge. This includes the reproducibility of calculations, the replication of findings with new evidence, and the transparency of decisions made during an investigation. The distinctions matter because a failure at one level does not identify a defect at every other level.

Simmons, Nelson, and Simonsohn's 2011 False-Positive Psychology examined how undisclosed flexibility in data collection and analysis could generate apparently persuasive results. Decisions about stopping, outcomes, exclusions, and analytical choices could accumulate without appearing in the final report. The diagnostic issue was the gap between the research process and its published representation. [77]

The Open Science Collaboration's 2015 project attempted replications of 100 studies from three psychology journals. Its findings prompted further examination of effect estimates, replication criteria, and research practices. The project concerns a particular collection of studies. It should not be converted into a numerical verdict on every field, method, or individual paper. [78]

Nosek and colleagues' 2018 account of preregistration explained how a time-stamped plan can separate advance predictions from analyses developed after seeing the evidence. Exploration remains valuable. The distinction allows readers to assess the evidential role of a result with a more accurate account of how it was produced. [79]

The National Academies' 2019 report distinguishes computational reproducibility using the same data and procedures from replicability using new data. Terminology has varied across communities, so the intended meaning should be stated. An analysis can be computationally reproducible and still rely on an invalid measure or an unsupported causal assumption. [80]

Preregistration likewise does not certify a good design. A researcher can commit in advance to a weak measure or an unsuitable model. Unexpected events may justify a change. The useful record preserves the original plan, the modification, its reason, and its timing relative to inspection of the outcomes.

A failed replication is a diagnostic starting point. Possible explanations include sampling variation, differences in implementation, changes in the population, an exaggerated original estimate, and defects in either investigation. The evidence should narrow these alternatives. Treating every discrepancy as misconduct would lose the opportunity to understand the actual failure mode.

Research institutions can preserve the artifacts needed for this work through data documentation, code, protocols, and appropriate access arrangements. The aim is not unrestricted release of every participant record. It is an arrangement in which relevant claims can be examined while obligations to participants remain effective. Reproducibility is therefore both a technical property of an analysis and an institutional achievement.

## 30. Algorithms and generative models as social research instruments

Algorithms can classify documents, estimate unobserved attributes, and support the examination of large collections of social evidence. Their output introduces another representation that needs validation. A model label is not automatically an observation of the social condition named by that label.

Obermeyer and colleagues' 2019 study of an algorithm used in population health management provides a specific example. Predicting expenditure as a proxy for health need contributed to racial bias because expenditure and need were related differently across groups. The defect could not be understood solely by asking whether the algorithm predicted its training target accurately. The diagnostic boundary included the choice of target. [81]

This problem extends to social research instruments. A sentiment classifier may perform well on its benchmark while misunderstanding irony or a community's vocabulary. An automated occupation code may reproduce historical categories whose meaning has changed. Evaluation should examine the intended population and use, as well as the average performance on a convenient dataset.

Argyle and colleagues' 2023 Out of One, Many examined GPT-3 as a means of simulating response distributions under demographic conditioning. The reported comparisons concern particular tasks, data, and model conditions. They do not establish that generated respondents are interchangeable with human participants for arbitrary questions or populations. [82]

Bail's 2024 discussion of generative AI in social science examined possible research uses together with their limitations. For our purposes, the important distinction is between a model producing a plausible social account and an investigation establishing a claim about people. A generated explanation can organize hypotheses. It does not become historical evidence merely because it is coherent. [83]

AI-assisted qualitative coding requires an inspectable relationship between source passage and assigned category. A useful record includes the model and prompt configuration, the coding scheme, the evidence span, and the human decisions involved in accepting or revising the result. Repeated agreement among outputs from one model does not constitute independent corroboration.

Synthetic data also requires careful interpretation. A generated dataset may be useful for testing code, protecting some forms of information, or exploring a model. It cannot reveal an unobserved social fact unless the inference connecting it to empirical evidence is justified. In particular, generating more rows from a model does not create more independent observations of the population.

The diagnostic opportunity is substantial because these systems also leave traces of their own operation. We can examine transformations, compare versions, locate systematic disagreements, and test changes. The resulting investigation must retain both levels: the behavior of the research instrument and the social process the instrument is intended to represent.

## PART VII

## Diagnostic cases, patterns, and the preservation of explanations

### 31. Polling failure and the reconstruction of an analysis

Well-known failures can become simplified stories. The simplification is itself a diagnostic problem when it replaces examination of the original evidence. Two cases illustrate the need to separate coverage, response, processing, and interpretation.

The Literary Digest's 1936 presidential poll incorrectly favored Alfred Landon over Franklin Roosevelt. The case is often explained only through a sampling frame that overrepresented more affluent people. Squire's 1988 reanalysis examined both the sample and response process and gave substantial attention to nonresponse. The historical lesson is not adequately preserved by the statement that the sample was simply too small or that one frame defect explains everything. [84]

The number of responses can make a poll look authoritative while leaving selection problems intact. If response is related to political preference in ways not captured by the adjustment information, a large response count can yield a precise description of the respondents and a poor description of the electorate. The relevant diagnostic artifact includes invitations and response patterns, not only the returned answers.

The debt-and-growth controversy examined by Herndon, Ash, and Pollin provides a different failure mode. Their critique appeared as a working paper in 2013 and as a journal article in 2014. It identified a spreadsheet coding error together with data-selection and weighting choices that affected the reported relationship in Reinhart and Rogoff's analysis. Treating the episode as only a spreadsheet mistake would omit consequential analytical choices. [85]

Here literal debugging and methodological criticism meet. A formula range can be inspected and corrected. A weighting scheme requires an explanation of which average the analysis intends to estimate. A selection rule requires a justification for including some observations and excluding others. These operations occur in one workflow but answer different questions.

Correcting the calculation does not by itself determine the causal effect of public debt on growth. A descriptive association can reflect reverse causation, common causes, and differences between historical settings. The correction changes what the analyzed data support; it does not settle every policy question associated with the topic.

Both cases show why the final published table is insufficient for a complete postmortem. The investigator needs the history of selection and transformation. The strongest correction identifies the defect, shows its effect on the result, and limits the revised conclusion to what the evidence establishes.

A memorable failure story should remain open to the same source criticism as any other historical account.

## **32. Mixed evidence and the investigation of disagreement**

A social investigation may combine interviews, observations, administrative records, surveys, and experiments. The value of the combination depends on how the sources relate to one another. Collecting different formats does not automatically create an integrated explanation.

Fetters, Curry, and Creswell's 2013 account of mixed-methods integration distinguished integration at the levels of design, methods, and interpretation. A qualitative phase can inform the construction of a survey, or qualitative evidence can help explain a quantitative result. Integration requires a relationship between the inquiries, not merely their placement in adjacent chapters of a report. [86]

Disagreement is often the most useful point of contact. Suppose a service survey reports satisfaction while interviews describe severe difficulties. The findings may concern different populations. The survey may ask about the courtesy of staff, while the interviews concern the outcome of the service. People may also express satisfaction relative to low expectations. Before declaring one source wrong, we should examine what each observation means.

Sources can agree for a shared reason that does not provide independent support. An interviewee may repeat a statistic from an agency report, and a newspaper may quote the same interview. Three artifacts then preserve one information path. The relationship resembles a trace set in which several records are copies of the same underlying message.

Conversely, different methods can provide complementary evidence without measuring the same thing. Administrative data may locate a delay. Observation may reveal how a handoff occurs. Interviews may explain why staff interpret the rule as they do. The combined account is useful because the artifacts constrain different parts of the explanation.

A diagnostic comparison should retain the original units. A household interview cannot be merged directly with an individual outcome without a rule connecting the two. An annual survey and daily transaction records need an account of their time windows. A qualitative category and a numerical indicator need an explanation of how they correspond.

We can therefore treat integration as an explicit mapping between evidence and claims. The mapping may reveal corroboration, complementarity, or conflict. Conflict should remain visible until an explanation is supported. Prematurely forcing agreement can remove the very observation that would have exposed a mistaken category or a missing mechanism.

## **33. Connections to pattern-oriented software diagnostics**

Our analysis patterns describe recurring structures in software traces, logs, narratives, and data. The pattern language helps an analyst recognize a useful relationship and choose a further examination. Its

application to social evidence is a transfer of analysis questions. It does not establish that social activity obeys the execution rules of a computer. [87, 88]

An Adjoint Thread of Activity groups messages through a selected invariant attribute or criterion. In a social record collection, we can similarly follow a case identifier across offices or follow a category of request across many people. Such a grouping is an analytical construction. It may combine records produced by different participants and should not be mistaken for one person's continuous experience.

Basic Facts provide another useful connection. Before explaining a discrepancy, we establish what the artifacts directly support: their dates, recorded statuses, definitions, and scope. An interpretation may later connect these facts through a proposed mechanism. Keeping that transition visible helps prevent a plausible narrative from acquiring the status of an observation.

The following mappings are proposed applications. The final column identifies a boundary that the transfer must preserve.

| Analysis pattern                  | Proposed application to social evidence                              | Boundary to preserve                                   |
|-----------------------------------|----------------------------------------------------------------------|--------------------------------------------------------|
| <b>Basic Facts</b>                | Establish recorded dates, categories, counts, and source identities. | A recorded assertion can still be inaccurate.          |
| <b>Adjoint Thread of Activity</b> | Follow one case, institution, or category across records.            | A grouping criterion does not create a single actor.   |
| <b>Incomplete History</b>         | Identify intervals or stages for which evidence is missing.          | Missing evidence is not evidence of inactivity.        |
| <b>Activity Disruption</b>        | Locate a break in a recurring sequence of institutional actions.     | Check whether the recording procedure changed.         |
| <b>Delay Dynamics</b>             | Compare the timing of applications, decisions, and responses.        | Stored timestamps may refer to different events.       |
| <b>Message Invariant</b>          | Examine stable wording in forms or routine reports.                  | Stable wording does not guarantee stable meaning.      |
| <b>Inter-Correlation</b>          | Compare changes across separate record collections.                  | Co-movement does not identify a causal mechanism.      |
| <b>Causal Chains</b>              | Organize a candidate sequence of enabling events.                    | Each proposed connection needs causal evidence.        |
| <b>CoTrace</b>                    | Read related traces or records together.                             | Shared provenance can limit independent corroboration. |
| <b>Data Association</b>           | Link records that concern a proposed common entity.                  | Incorrect linkage can manufacture a relationship.      |

The strongest use of these patterns is to direct attention. A break suggests checking acquisition and institutional change. A delay suggests inspecting event definitions and transitions. A repeated message suggests examining whether a template conceals important differences. The pattern does not provide the substantive social explanation in advance.

This distinction also protects the historical account. We can recognize a relationship between an older inquiry and a later pattern without claiming influence. A nineteenth-century investigator did not need our vocabulary to compare records or question a classification. The connection concerns a recurring diagnostic problem, while the historical method retains its own setting and purpose.

## 34. A worked example of a misleading improvement

Consider a hypothetical public service that reports an improvement in its completion rate. The following figures are constructed to illustrate an investigation. They are not observations from a real institution. Each period contains 100 initial requests, and the underlying service outcomes are deliberately held equal.

| Recorded outcome by the reporting date | Earlier period | Later period |
|----------------------------------------|----------------|--------------|
| Initial requests                       | 100            | 100          |
| Service delivered and verified         | 60             | 60           |
| Closed without service delivery        | 20             | 20           |
| Still open                             | 20             | 20           |
| Published completion indicator         | 60 percent     | 80 percent   |

The first task is to state the discrepancy. The published indicator rises, while the number of verified service deliveries remains unchanged. The question is whether the indicator measures the same outcome in both periods. This is more precise than beginning with the general statement that the service data are unreliable.

The analyst obtains the reporting definitions and finds a change. The earlier indicator counted verified delivery. The later indicator counted both verified delivery and closure without delivery. The numerator changed from 60 to 80, while the denominator remained 100. The arithmetic is correct under each rule. The historical comparison is misleading if both values are described as the same completion rate.

Following request identifiers across the records provides a second examination. The selected Adjoint Threads of Activity distinguish application, acknowledgment, delivery, closure, and reopening. Cases closed without delivery should remain identifiable as such. If the reporting extract contains only the latest status, the earlier sequence may have been lost.

Participant interviews can answer an additional question: what did closure mean for people seeking the service? They cannot change the count in the constructed table, but they can reveal consequences omitted from the indicator. Closure may end an administrative workflow while leaving the original need unresolved. The service's purpose therefore matters to the interpretation of success.

The immediate correction is to preserve separate indicators for verified delivery and administrative closure, with explicit definitions. Historical values can be recalculated under a common rule if the underlying records permit it. The original published series should remain identifiable, together with the reason for revision. This retains the evidence needed to understand the discrepancy.

The example establishes a measurement discontinuity, not a causal evaluation of the service. It does not show why 20 requests remained open or whether an intervention would improve delivery. Those questions require further evidence. A useful diagnostic conclusion states what was found and leaves the unresolved questions at their proper level.

## 35. Diagnostic memory and the limits of reconstruction

An investigation becomes more valuable when its evidence and corrections remain available for later examination. This requires more than retaining a final report. The history of the dataset, the definitions used at each stage, the code, and the reasons for revisions can all affect what a later analyst can reconstruct.

The developments reviewed here add different forms of diagnostic memory. A field notebook preserves context. A sampling record preserves selection. A preregistration preserves the timing of commitments. A versioned script preserves transformations. A participant's correction preserves a challenge to the original account. These artifacts should be connected without erasing their different evidential roles.

Digital preservation also has limits. An archive can retain data while losing the software needed to interpret it. A dataset can retain variable names while losing the coding manual. A model can be available while its original training or evaluation conditions are not. The apparent survival of an artifact does not ensure the survival of the relationships that gave it meaning. The reproducibility questions discussed by the National Academies therefore extend into the organization of research memory. [80]

Some histories remain underdetermined even when every surviving record has been examined. Several explanations may be compatible with the evidence. Further computation cannot supply an observation that was never made. The appropriate result can be a set of constrained alternatives, with an account of what additional evidence would distinguish them.

There are also limits to intervention. An institution can adapt to a correction, participants can change their interpretation of a measure, and a successful local arrangement can alter the surrounding environment. A diagnostic account needs a scope and a time. It should not silently turn a supported explanation of one setting into a permanent description of all similar systems.

Automation can assist by comparing versions, locating inconsistent codes, and organizing large evidence collections. Its contribution is strongest when the relationship between source and result remains inspectable. Automated output that obscures provenance can reduce the very diagnostic capacity it was intended to improve.

The continuing task is to preserve an explanation together with the evidence that supports it and the alternatives it has not excluded. This is a practical requirement for cumulative social inquiry. It allows a later investigator to distinguish a corrected mistake, a changed social process, and a different question being asked of the same surviving artifacts.

## Conclusion

The history of diagnostics in social sciences is a history of relationships between social activity, its representation, and its explanation. Political arithmetic made population regularities available for comparison. Surveys and fieldwork constructed new records of social conditions. Measurement theory examined the interpretation of indicators. Experimental and causal methods made comparisons and assumptions more explicit. Institutional and participatory approaches examined the setting and authority of intervention. Computational methods added new traces and new instruments requiring their own diagnosis.

These developments did not replace one another in a simple sequence. A contemporary algorithm may require a careful interview to reveal that its target has the wrong meaning. A historical record may require a numerical check to expose an implausible claim. A randomized study may need an account of implementation to explain why its result differs across settings. Diagnostic skill includes choosing the level at which a question can be answered and moving between levels when the evidence requires it.

Our pattern-oriented perspective provides a language for some recurring operations: establishing facts, reconstructing activities, comparing artifacts, examining gaps, and locating changes in representation. The language is useful when it directs attention to evidence. It becomes misleading if a familiar pattern is treated as sufficient proof of a social mechanism.

We can therefore understand social diagnostic progress as an increasing capacity to examine the production of explanations. The useful result is an account that says what was observed, how the observations were transformed, which inference is supported, and where the evidence ends. Such an account remains open to correction because it preserves the material through which a correction can be made.

## Appendix A. Selected chronology

The dates identify the stated publication, project, or observation period. The final column summarizes its relevance to this history. It does not assign exclusive priority for a method or imply a direct line of influence between adjacent entries.

| Date      | Work or development                                                          | Diagnostic relevance                                                            |
|-----------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| 1377      | Drafting of Ibn Khaldun's Muqaddimah [3]                                     | Criticism of reports through the conditions of social organization.             |
| 1662      | Graunt's observations on Bills of Mortality [4]                              | Repeated records become evidence about population patterns.                     |
| 1690      | Posthumous publication of Petty's Political Arithmetick [5]                  | Numerical estimates support political and economic comparison.                  |
| 1833      | Guerry's moral statistics of France [6]                                      | Tables and maps expose aggregate social distributions.                          |
| 1835      | Quetelet's work on social physics [7]                                        | Statistical regularity becomes an object of social explanation.                 |
| 1838      | Martineau's How to Observe [8]                                               | Observation and interpretation require explicit discipline.                     |
| 1843      | First edition of Mill's System of Logic [40]                                 | Comparative inquiry receives an explicit logical treatment.                     |
| 1845      | Engels's study of the English working class [12]                             | Social conditions are examined through structural relationships.                |
| 1867      | First volume of Marx's Capital [13]                                          | Administrative reports enter a structural criticism of economic relationships.  |
| 1883      | Galton's Inquiries into Human Faculty [23]                                   | Measurement develops alongside eugenic purposes requiring historical criticism. |
| 1886–1903 | Booth's London inquiry [9]                                                   | Urban observation is connected with maps and classifications.                   |
| 1895      | Hull-House Maps and Papers and Durkheim's Rules [10, 14]                     | Field evidence and methodological argument develop in different settings.       |
| 1897      | Durkheim's Suicide [15]                                                      | Recorded rates support comparative social explanation.                          |
| 1899      | Du Bois's The Philadelphia Negro [11]                                        | Community investigation joins historical and empirical evidence.                |
| 1904      | Weber on objectivity and Spearman on intelligence [16, 24]                   | Concepts and latent structure become explicit analytical problems.              |
| 1917      | Richmond's Social Diagnosis [1]                                              | Social evidence is systematically examined for casework.                        |
| 1922      | Malinowski's Argonauts [17]                                                  | Fieldwork records preserve activity and context.                                |
| 1932      | Likert's attitude-measurement monograph [25]                                 | Sets of responses become instruments for measuring attitudes.                   |
| 1934      | Neyman on sampling, Kuznets on income, Moreno on interrelations [20, 43, 49] | Population selection, economic aggregation, and relational evidence develop.    |
| 1935      | Fisher's Design of Experiments [31]                                          | Randomization and replication support designed comparisons.                     |

| Date          | Work or development                                                                                                          | Diagnostic relevance                                                                       |
|---------------|------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| 1936          | Literary Digest polling failure, reexamined in 1988 [84]                                                                     | Coverage and response require separate investigation.                                      |
| 1944          | Haavelmo on econometrics and Mahalanobis on surveys [33, 21]                                                                 | Models and acquisition designs are connected with probabilistic inference.                 |
| 1946–1948     | Lewin on action research and Merton on prophecy [57, 62]                                                                     | Intervention and expectations become part of the observed process.                         |
| 1950          | Robinson on ecological correlations [51]                                                                                     | Aggregate and individual relationships are distinguished.                                  |
| 1955–1959     | Construct validation and the multitrait-multimethod matrix [26, 27]                                                          | Interpretation of a measure becomes a sustained evidential problem.                        |
| 1960–1963     | Regression discontinuity and design taxonomies [37, 32]                                                                      | Comparison designs are examined through their identifying conditions.                      |
| 1967          | Grounded theory and ethnomethodology texts [19, 53]                                                                          | Category formation and the production of social order are examined.                        |
| 1968          | Panel Study of Income Dynamics begins [46]                                                                                   | Linked observations support histories of households and individuals.                       |
| 1970          | English publication of Freire's Pedagogy of the Oppressed [58]                                                               | Dialogue and authority over knowledge enter the account of inquiry.                        |
| 1971–1974     | Schelling, Cox, Granovetter, Geertz, Rubin, and conversation analysis [72, 47, 50, 18, 34, 54]                               | Mechanisms, event timing, relationships, meanings, and sequences gain new representations. |
| 1976–1979     | Lucas, Rubin on missingness, Cook, and Campbell [63, 75, 74, 64]                                                             | Stability, absence, influence, and target pressure become diagnostic questions.            |
| 1979          | Belmont Report [65]                                                                                                          | Research procedures are connected with obligations to participants.                        |
| 1983–1989     | Propensity scores, Normal Accidents, comparative methods, and latent-variable synthesis [35, 55, 41, 28]                     | Causal, organizational, configurational, and measurement assumptions are made explicit.    |
| 1990–1991     | Human Development Report, Governing the Commons, and Action and Knowledge [44, 60, 59]                                       | Development, institutions, and participation broaden evaluative inquiry.                   |
| 1993–1996     | Invariance, policy comparison, spatial and sequence methods, instrumental variables, and Sugarscape [29, 38, 52, 48, 39, 73] | Comparability and the relationship between local evidence and general claims are examined. |
| 1997–1999     | Realistic Evaluation, Seeing Like a State, Decolonizing Methodologies, and Development as Freedom [56, 61, 66, 45]           | Context, local knowledge, authority, and substantive purposes become central.              |
| 2000 and 2009 | First and second editions of Pearl's Causality [36]                                                                          | Graphical and structural assumptions support explicit causal questions.                    |
| 2007          | Ostrom's diagnostic approach beyond panaceas [2]                                                                             | Institutional analysis is organized as contextual diagnosis.                               |
| 2009–2010     | Computational social science and the WEIRD critique [69, 30]                                                                 | New data scale is accompanied by renewed questions about population reach.                 |
| 2011–2014     | Research flexibility, process tracing, mixed methods, and prominent reanalyses [77, 42, 86, 85, 71]                          | Evidence production and analytical decisions become objects of inspection.                 |

| Date      | Work or development                                                                                       | Diagnostic relevance                                                              |
|-----------|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| 2015      | Psychology replication project and mobile-metadata poverty prediction [78, 70]                            | Repeated studies and proxy-based prediction expose different validation problems. |
| 2018–2019 | Preregistration, many-analyst comparison, reproducibility synthesis, and algorithm audit [79, 76, 80, 81] | Research histories and target definitions become more inspectable.                |
| 2020      | CARE Principles and Data Feminism [67, 68]                                                                | Data governance, context, and power enter the evidential account.                 |
| 2023–2024 | Language-model samples and assessment of generative AI in social research [82, 83]                        | Generated social representations require validation against human evidence.       |

## Appendix B. Glossary

**Abduction.** Reasoning that proposes an explanation for observed evidence. A plausible explanation remains a candidate until it is examined against alternatives and further observations.

**Administrative data.** Records produced in operating an institution or service. Their categories reflect operational purposes, which may differ from the purposes of a later research study.

**Adjoint Thread of Activity.** In our analysis-pattern language, a sequence selected through a shared attribute or criterion. A proposed social application might follow one case across several institutions.

**Audit trail.** A preserved history of actions and transformations that allows a result to be connected with its inputs and decisions.

**Causal effect.** A difference between outcomes under specified intervention alternatives. The population, intervention, outcome, and time horizon form part of its definition.

**Causal graph.** A graphical representation of assumed causal relationships. It supports inferences under those assumptions and does not become correct merely by being drawn.

**Censoring.** Partial observation of an event time, for example when observation ends before the event occurs. Its treatment depends on assumptions about the observation process.

**Confounding.** A situation in which the comparison used to estimate a causal effect also reflects other causes associated with treatment or exposure.

**Construct.** A concept used to organize explanation, such as trust or perceived control. Indicators provide evidence about it under an interpretation that requires support.

**Construct validity.** The evidential support for interpreting a measure in terms of a proposed construct and its theoretical relationships.

**Coverage error.** A mismatch between the intended population and the population accessible through the sampling or recording system.

**Diagnostic artifact.** A preserved representation used to support, question, or refine an explanation. Its production and transformation are part of its evidential meaning.

**Ecological inference.** Inference involving aggregate data and relationships at another level, often individuals. Directly transferring an aggregate association can produce an ecological fallacy.

**Estimand.** The quantity a study intends to learn. It should be distinguished from the procedure used to estimate it and from the numerical estimate obtained.

**Event history.** A representation of events, transitions, and periods at risk over time. The definitions of event and time origin affect the analysis.

**External validity.** Support for extending an inference beyond the particular study conditions. Its scope should be stated rather than treated as an unrestricted property.

**Identifiability.** The extent to which a target quantity or explanation is determined by the assumed model and available observations. Additional data of the same kind may not resolve nonidentifiability.

**Indicator.** An observable quantity used to represent a condition or construct. Its usefulness depends on the intended interpretation and the process that generates it.

**Instrumental variable.** A variable used to identify a causal effect through treatment variation under specified assumptions. Relevance alone does not establish a valid instrument.

**Internal validity.** The support for the causal interpretation of a result within the study's conditions, including the exclusion of relevant rival explanations.

**Latent variable.** A variable not directly observed but represented through a model relating it to indicators or other variables.

**Measurement invariance.** Conditions under which a measurement model supports specified comparisons across groups or occasions. Different comparisons require different forms of invariance.

**Mechanism.** A proposed process through which conditions and activities produce an outcome. Naming a mechanism does not establish that it operated in the case.

**Missing at random.** A condition under which missingness may depend on observed information but not on the missing values after conditioning on the observed information in the model.

**Nonresponse.** Failure to obtain an intended response from a selected unit or to a particular item. Its consequences depend on how it relates to the target quantity.

**Observability.** In a specified dynamical model, a property concerning reconstruction of internal states from available outputs. General visibility of social records is a broader and weaker notion.

**Panel.** A study that repeatedly observes linked units over time. Attrition and changes in unit definition affect the histories that can be reconstructed.

**Potential outcome.** The outcome associated with a unit under a specified treatment or intervention condition, whether or not that condition is realized.

**Preregistration.** A time-stamped record of planned research decisions. It helps distinguish advance commitments from decisions made after inspecting the relevant evidence.

**Process tracing.** Examination of within-case evidence in relation to a proposed causal process and rival explanations. A chronology is one input, not a sufficient causal argument.

**Proxy.** An observed measure used in place of a target that is harder to observe. A predictive association does not ensure suitability for every use.

**Quasi-experiment.** A study using structured variation without investigator-controlled random assignment to support causal inference under explicit conditions.

**Reactivity.** A change in behavior associated with measurement, observation, or participation in a study. Its presence and mechanism require evidence in the particular setting.

**Reflexivity.** The relationship through which knowledge, categories, or investigation can affect the people and processes being studied; also critical examination of the investigator's own position.

**Reliability.** Consistency of measurement under specified conditions. It does not establish that the resulting interpretation is valid.

**Replicability.** In the terminology used here, consistency of findings examined with newly obtained data. Other communities sometimes use the term differently.

**Reproducibility.** In the terminology used here, the ability to obtain consistent computational results using the same data, methods, and analytical conditions.

**Sampling frame.** The operational representation from which units are selected. It may omit eligible units or include units outside the target population.

**Sensitivity analysis.** Examination of how conclusions change when assumptions, specifications, or other choices are varied within a stated range.

**Social diagnosis.** Investigation of social conditions and relationships for an explanatory or practical purpose. Its historical uses include social casework and contextual institutional analysis.

**Sociogram.** A diagram of selected relationships among people or groups. Its meaning depends on how the relationships and boundaries were defined.

**Thick description.** An interpretive account that situates an action within the meanings and conventions that make it intelligible.

**Total survey error.** A framework for considering multiple sources of survey error, including coverage, sampling, nonresponse, measurement, and processing.

**Transportability.** The support, under explicit assumptions, for applying a causal inference from one population or setting to another.

**Triangulation.** Examination of a claim through multiple forms or sources of evidence. Independence, comparability, and disagreement must be assessed rather than assumed.

**Validity.** Evidential support for a specified interpretation, inference, or use. A measure or method is not valid for all purposes simply because one use is supported.

## Appendix C. Evidence taxonomy and blind spots

This taxonomy classifies artifacts by the relationships they can make available for analysis. It is not a universal ranking of methods. A source that is weak for estimating prevalence may be essential for interpreting an action, while a precise population estimate may leave its mechanism unresolved.

| Artifact                                   | What it can support                                                  | Characteristic blind spot                                                        |
|--------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------------------|
| <b>Register or administrative file</b>     | Recorded events, statuses, and institutional decisions.              | Unrecorded activity and changes in operational definitions.                      |
| <b>Census or population table</b>          | Distributions within a stated enumeration and classification.        | Coverage differences and relationships hidden by aggregation.                    |
| <b>Questionnaire and response file</b>     | Answers produced under a defined invitation and question sequence.   | Excluded people, nonresponse, question interpretation, and response incentives.  |
| <b>Interview recording</b>                 | Situated accounts, explanations, and interactions.                   | Selective memory, strategic presentation, and the interview relationship.        |
| <b>Field notes</b>                         | Observed practices and contextual detail.                            | Selective access and the observer's categories and attention.                    |
| <b>Map or spatial extract</b>              | Locations and spatial distributions under specified boundaries.      | Boundary effects, exposure differences, and institutional recording intensity.   |
| <b>Network data</b>                        | Specified relationships among observed units.                        | Missing ties, omitted nodes, and ambiguity in edge meaning.                      |
| <b>Panel or event history</b>              | Linked transitions and durations over time.                          | Attrition, censoring, uncertain dates, and changed unit definitions.             |
| <b>Experimental record</b>                 | Assignment, delivery, and outcome comparisons.                       | Noncompliance, spillovers, measurement defects, and limited transfer.            |
| <b>Historical document collection</b>      | Evidence about sequences, decisions, and public or private accounts. | Selective survival, dependence between sources, and retrospective justification. |
| <b>Statistical model output</b>            | Relationships implied or estimated under a specification.            | Nonidentifiability, misspecification, and weak substantive interpretation.       |
| <b>Simulation trace</b>                    | Consequences of implemented rules and initial conditions.            | Equating model behavior with an established empirical mechanism.                 |
| <b>Digital platform trace</b>              | Activity visible through a particular software infrastructure.       | Platform selection, interface effects, filtering, and undocumented change.       |
| <b>Code and computational log</b>          | Data transformations and execution behavior.                         | Correct execution of an inappropriate analytical plan.                           |
| <b>Preregistration and version history</b> | Timing of plans, modifications, and analytical commitments.          | Advance commitment to a poor design or incomplete recording of deviations.       |
| <b>Generated or AI-coded output</b>        | Model-produced representations under specified conditions.           | Unsupported factual content, hidden dependencies, and a mistaken target.         |

## **C1 Questions to carry with an artifact**

The first question concerns origin. What activity produced the artifact, and what purpose did that activity serve? A record produced to authorize payment has a different relationship to behavior from a record produced for a confidential interview. The original purpose can explain both the detail that survives and the detail that is absent.

The second question concerns selection. Which people, events, places, or periods could enter the artifact? Which were excluded, lost, or not recorded? The answer defines the population about which the artifact can directly inform us. An analyst may extend the inference, but that extension needs its own justification.

The third question concerns transformation. How did the original record become a code, variable, summary, or model input? The history should preserve the definitions and operations through which information was added, combined, or discarded. A later correction is easier when the earlier state remains identifiable.

The fourth question concerns the claim. Does the artifact support a description, a predictive relationship, a causal comparison, an interpretation, or an evaluation? A source can be strong for one of these purposes and weak for another. The statement of the claim should not exceed the relationship the evidence establishes.

The fifth question concerns challenge and consequence. Who can inspect or correct the record, and what decisions depend on it? These relationships affect the production of evidence and the costs of error. They also determine whether a diagnostic finding can lead to an effective and accountable correction.

## **C2 Combining evidence without losing its meaning**

We begin with the problem description and identify the inferential step that remains uncertain. Additional evidence is useful when it bears on that step. If the uncertainty concerns population coverage, another analysis of the same respondents may contribute little. If it concerns the meaning of a recorded status, a coding manual or participant account may be more informative than a larger sample.

The combined account should retain independent paths from source to conclusion. Agreement among sources is stronger when their relevant failure modes differ. Where they share a classification system, reporting incentive, or original informant, agreement may preserve the same limitation several times.

Finally, we preserve unresolved alternatives. A diagnostic report can establish that an indicator changed definition while leaving the underlying social trend uncertain. This is a useful result because it prevents a misleading inference and identifies the evidence still required. The completeness of an explanation should be judged by the support for its claims, not by the absence of visible uncertainty.

## References

References are numbered by first appearance. Original publication years are retained where later translations, reissues, or digital copies are linked. Web resources were consulted on September 28–29, 2026.

- [1] Richmond, Mary E. 1917. *Social Diagnosis*. New York: Russell Sage Foundation.  
<https://www.russellsage.org/publications/book/social-diagnosis>
- [2] Ostrom, Elinor. 2007. 'A Diagnostic Approach for Going beyond Panaceas.' *Proceedings of the National Academy of Sciences* 104(39): 15181–15187.  
<https://doi.org/10.1073/pnas.0702288104>
- [3] Ibn Khaldun. 1958. *The Muqaddimah: An Introduction to History*. Translated by Franz Rosenthal. 3 volumes. Bollingen Series XLIII. New York: Pantheon Books. Original Arabic work drafted in 1377.  
<https://search.worldcat.org/title/muqaddimah-an-introduction-to-history/oclc/307867>
- [4] Graunt, John. 1662. *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*. London: printed by Thomas Roycroft for John Martin, James Allestry, and Thomas Dicus.  
<https://wellcomecollection.org/works/ear22f7e>
- [5] Petty, William. 1690. *Political Arithmetick*. London: Robert Clavel and Henry Mortlock. Published posthumously.  
<https://www.marxists.org/reference/subject/economics/petty/>
- [6] Guerry, André-Michel. 1833. *Essai sur la statistique morale de la France*. Paris: Crochard.  
<https://gallica.bnf.fr/ark:/12148/bpt6k842782>
- [7] Quetelet, Adolphe. 1835. *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*. 2 volumes. Paris: Bachelier.  
<https://wellcomecollection.org/works/aebc5kja>
- [8] Martineau, Harriet. 1838. *How to Observe: Morals and Manners*. London: Charles Knight. Digitized text at Project Gutenberg.  
<https://www.gutenberg.org/ebooks/33944>
- [9] London School of Economics and Political Science Library. Charles Booth's London: What Were the Poverty Maps? Archival guide to the inquiry of 1886–1903, distinguishing the earlier maps from the 1898–1899 survey maps published in 1902–1903.  
<https://booth.lse.ac.uk/learn-more/what-were-the-poverty-maps>
- [10] Residents of Hull-House. 1895. *Hull-House Maps and Papers*. New York: Thomas Y. Crowell. Original map material and publication metadata preserved by Cornell University Library.  
<https://digital.library.cornell.edu/catalog/ss:3293796>
- [11] Du Bois, W. E. B. 1899. *The Philadelphia Negro: A Social Study*. Philadelphia: University of Pennsylvania. Includes Isabel Eaton's study of domestic service.  
<https://www.loc.gov/item/05033530/>
- [12] Engels, Friedrich. 1845. *The Condition of the Working Class in England*. Originally published in German. Authorized English edition first published in New York in 1887; linked text from the 1969 Panther edition.  
<https://www.marxists.org/archive/marx/works/1845/condition-working-class/>
- [13] Marx, Karl. 1867. *Capital*, Volume I. Originally published in German. Chapter on the working day, numbered 10 in the linked English text.  
<https://www.marxists.org/archive/marx/works/1867-c1/ch10.htm>
- [14] Durkheim, Émile. 1895. *Les Règles de la méthode sociologique*. Paris: Félix Alcan. English extracts in *Classical Sociological Theory and Foundations of American Sociology*.  
<https://open.oregonstate.education/sociologicaltheory/chapter/rules-of-method/>

- [15] Durkheim, Émile. 1897. *Le Suicide: Étude de sociologie*. Paris: Félix Alcan. Bibliographic record and editions in The Durkheim Pages, University of Chicago.  
<https://durkheim.uchicago.edu/Bibliography/Bib01.html>
- [16] Weber, Max. 1904. 'Objectivity' in *Social Science and Social Policy*. Reprinted in *The Methodology of the Social Sciences*, translated and edited by Edward A. Shils and Henry A. Finch, Free Press, 1949. Linked extracts from Wolf Heydebrand (ed.), *Max Weber: Sociological Writings*, Continuum, 1994.  
<https://www.d.umn.edu/cla/faculty/jhamlin/4111/Readings/Science%20and%20Sociology%20by%20Max%20Weber.htm>
- [17] Malinowski, Bronislaw. 1922. *Argonauts of the Western Pacific*. London: George Routledge & Sons. Introduction reproduced in *An Open Introduction to Anthropology*, University of Lethbridge.  
<https://pressbooks.openeducationalberta.ca/anth/chapter/argonauts-of-the-western-pacific-introduction/>
- [18] Geertz, Clifford. 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books. Especially 'Thick Description: Toward an Interpretive Theory of Culture.'  
<https://search.worldcat.org/oclc/737285>
- [19] Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.  
<https://books.google.com/books?id=oUxEQAIAAJ>
- [20] Neyman, Jerzy. 1934. 'On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection.' *Journal of the Royal Statistical Society* 97(4): 558–625 (including discussion).  
<https://doi.org/10.2307/2342192>
- [21] Mahalanobis, Prasanta Chandra. 1944. 'On Large-Scale Sample Surveys.' *Philosophical Transactions of the Royal Society of London, Series B* 231(584): 329–451.  
<https://doi.org/10.1098/rstb.1944.0002>
- [22] Groves, Robert M., and Lars Lyberg. 2010. 'Total Survey Error: Past, Present, and Future.' *Public Opinion Quarterly* 74(5): 849–879.  
<https://doi.org/10.1093/poq/nfq065>
- [23] Galton, Francis. 1883. *Inquiries into Human Faculty and Its Development*. London: Macmillan. Linked archive includes the later second-edition text.  
<https://galton.org/books/human-faculty/>
- [24] Spearman, Charles. 1904. "'General Intelligence," Objectively Determined and Measured.' *American Journal of Psychology* 15(2): 201–292. Original article reproduced by *Classics in the History of Psychology*, York University.  
<https://psychclassics.yorku.ca/Spearman/>
- [25] Likert, Rensis. 1932. *A Technique for the Measurement of Attitudes*. *Archives of Psychology*, no. 140.  
<https://books.google.com/books?id=9rotAAAAYAAJ>
- [26] Cronbach, Lee J., and Paul E. Meehl. 1955. 'Construct Validity in Psychological Tests.' *Psychological Bulletin* 52(4): 281–302.  
<https://doi.org/10.1037/h0040957>
- [27] Campbell, Donald T., and Donald W. Fiske. 1959. 'Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.' *Psychological Bulletin* 56(2): 81–105.  
<https://doi.org/10.1037/h0046016>
- [28] Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.  
<https://doi.org/10.1002/9781118619179>
- [29] Meredith, William. 1993. 'Measurement Invariance, Factor Analysis and Factorial Invariance.' *Psychometrika* 58(4): 525–543.  
<https://doi.org/10.1007/BF02294825>
- [30] Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. 'The Weirdest People in the World?' *Behavioral and Brain Sciences* 33(2–3): 61–83, with subsequent commentary and response.  
<https://doi.org/10.1017/S0140525X0999152X>

- [31] Fisher, Ronald A. 1935. *The Design of Experiments*. First edition. Edinburgh: Oliver and Boyd.  
<https://search.worldcat.org/title/The-design-of-experiments/oclc/2417943>
- [32] Campbell, Donald T., and Julian C. Stanley. 1963. 'Experimental and Quasi-Experimental Designs for Research on Teaching.' In N. L. Gage (ed.), *Handbook of Research on Teaching*, 171–246. Chicago: Rand McNally. Linked copy is a later standalone reprint under the shorter title *Experimental and Quasi-Experimental Designs for Research*.  
<https://www.jameslindlibrary.org/campbell-dt-stanley-jc-1963/>
- [33] Haavelmo, Trygve. 1944. 'The Probability Approach in Econometrics.' *Econometrica* 12, supplement: iii–vi and 1–115.  
<https://doi.org/10.2307/1906935>
- [34] Rubin, Donald B. 1974. 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.' *Journal of Educational Psychology* 66(5): 688–701.  
<https://doi.org/10.1037/h0037350>
- [35] Rosenbaum, Paul R., and Donald B. Rubin. 1983. 'The Central Role of the Propensity Score in Observational Studies for Causal Effects.' *Biometrika* 70(1): 41–55.  
<https://doi.org/10.1093/biomet/70.1.41>
- [36] Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. Second edition. Cambridge University Press. First edition published in 2000.  
<https://bayes.cs.ucla.edu/BOOK-2K/>
- [37] Thistlethwaite, Donald L., and Donald T. Campbell. 1960. 'Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment.' *Journal of Educational Psychology* 51(6): 309–317.  
<https://doi.org/10.1037/h0044319>
- [38] Card, David, and Alan B. Krueger. 1994. 'Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania.' *American Economic Review* 84(4): 772–793. Linked NBER working paper appeared in 1993 and supplies the published citation.  
<https://www.nber.org/papers/w4509>
- [39] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. 'Identification of Causal Effects Using Instrumental Variables.' *Journal of the American Statistical Association* 91(434): 444–455.  
<https://doi.org/10.1080/01621459.1996.10476902>
- [40] Mill, John Stuart. 1843. *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker. Books III and VI. Linked text: eighth edition, New York: Harper & Brothers, 1882.  
<https://www.gutenberg.org/ebooks/27942>
- [41] Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California Press. Linked edition: 2014 reissue with a new introduction.  
<https://www.ucpress.edu/books/the-comparative-method/paper>
- [42] Collier, David. 2011. 'Understanding Process Tracing.' *PS: Political Science & Politics* 44(4): 823–830.  
<https://doi.org/10.1017/S1049096511001429>
- [43] Kuznets, Simon. 1934. *National Income, 1929–1932*. NBER Bulletin 49 (June 7). New York: National Bureau of Economic Research.  
<https://www.nber.org/books-and-chapters/national-income-1929-1932/national-income-1929-1932>
- [44] United Nations Development Programme. 1990. *Human Development Report 1990: Concept and Measurement of Human Development*. New York: Oxford University Press for UNDP.  
<https://hdr.undp.org/content/human-development-report-1990>
- [45] Sen, Amartya. 1999. *Development as Freedom*. New York: Alfred A. Knopf.  
<https://sen.scholars.harvard.edu/publications/development-freedom>
- [46] Panel Study of Income Dynamics. About the Panel Study of Income Dynamics Repository. University of Michigan and ICPSR. Institutional history of the study begun in 1968.  
<https://www.icpsr.umich.edu/sites/psid/home>

- [47] Cox, David R. 1972. 'Regression Models and Life-Tables.' *Journal of the Royal Statistical Society, Series B* 34(2): 187–202; discussion and reply extend through page 220.  
<https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>
- [48] Abbott, Andrew. 1995. 'Sequence Analysis: New Methods for Old Ideas.' *Annual Review of Sociology* 21: 93–113.  
<https://doi.org/10.1146/annurev.so.21.080195.000521>
- [49] Moreno, J. L. 1934. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Washington, DC: Nervous and Mental Disease Publishing Company.  
<https://wellcomecollection.org/works/kwjbgxeh>
- [50] Granovetter, Mark S. 1973. 'The Strength of Weak Ties.' *American Journal of Sociology* 78(6): 1360–1380.  
<https://doi.org/10.1086/225469>
- [51] Robinson, W. S. 1950. 'Ecological Correlations and the Behavior of Individuals.' *American Sociological Review* 15(3): 351–357.  
<https://www.jstor.org/stable/2087176>
- [52] Anselin, Luc. 1995. 'Local Indicators of Spatial Association—LISA.' *Geographical Analysis* 27(2): 93–115.  
<https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>
- [53] Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall. Publisher page for the later Polity reissue.  
<https://www.wiley-vch.de/en/areas-interest/humanities-social-sciences/studies-in-ethnomethodology-978-0-7456-0005-5>
- [54] Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. 'A Simplest Systematics for the Organization of Turn-Taking for Conversation.' *Language* 50(4): 696–735.  
<https://doi.org/10.2307/412243>
- [55] Perrow, Charles. 1984. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books.  
<https://books.google.com/books?id=N3hRAAAAMAAJ>
- [56] Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. London: Sage.  
<https://www.sagepub.com/shop/buy-a-book/realistic-evaluation-1-205276>
- [57] Lewin, Kurt. 1946. 'Action Research and Minority Problems.' *Journal of Social Issues* 2(4): 34–46.  
<https://doi.org/10.1111/j.1540-4560.1946.tb02295.x>
- [58] Freire, Paulo. 1970. *Pedagogy of the Oppressed*. Translated by Myra Bergman Ramos. New York: Herder and Herder. English publication of the Portuguese manuscript.  
<https://eric.ed.gov/?id=ED045793>
- [59] Fals-Borda, Orlando, and Muhammad Anisur Rahman, editors. 1991. *Action and Knowledge: Breaking the Monopoly with Participatory Action-Research*. New York: Apex Press; London: Intermediate Technology Publications.  
<https://practicalactionpublishing.com/book/20/action-and-knowledge>
- [60] Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. Design principles and teaching materials maintained by the Ostrom Workshop, Indiana University.  
<https://www.ostrom.indiana.edu/courses-teaching/teaching-tools/ostrom-design/index.html>
- [61] Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. Linked edition: Veritas paperback, 2020.  
<https://www.jstor.org/stable/j.ctvxkn7ds>
- [62] Merton, Robert K. 1948. 'The Self-Fulfilling Prophecy.' *The Antioch Review* 8(2): 193–210.  
<https://doi.org/10.2307/4609267>
- [63] Lucas, Robert E., Jr. 1976. 'Econometric Policy Evaluation: A Critique.' *Carnegie-Rochester Conference Series on Public Policy* 1: 19–46.  
[https://doi.org/10.1016/S0167-2231\(76\)80003-6](https://doi.org/10.1016/S0167-2231(76)80003-6)

- [64] Campbell, Donald T. 1979. 'Assessing the Impact of Planned Social Change.' *Evaluation and Program Planning* 2(1): 67–90. Earlier version: Dartmouth College Public Affairs Center, Occasional Paper 8, December 1976 (ERIC ED303512). [https://doi.org/10.1016/0149-7189\(79\)90048-X](https://doi.org/10.1016/0149-7189(79)90048-X)
- [65] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Completed September 30, 1978; published in the Federal Register April 18, 1979 (44 FR 23192). <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html>
- [66] Smith, Linda Tuhiwai. 1999. *Decolonizing Methodologies: Research and Indigenous Peoples*. London: Zed Books; Dunedin: University of Otago Press. <https://books.google.com/books?id=RK2KSQAACAAJ>
- [67] Carroll, Stephanie Russo, et al. 2020. 'The CARE Principles for Indigenous Data Governance.' *Data Science Journal* 19: 43. <https://doi.org/10.5334/dsj-2020-043>
- [68] D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. MIT Press. <https://data-feminism.mitpress.mit.edu/>
- [69] Lazer, David, et al. 2009. 'Computational Social Science.' *Science* 323(5915): 721–723. Archived manuscript title: 'Life in the network: the coming age of computational social science' (PMCID: PMC2745217). <https://doi.org/10.1126/science.1167742>
- [70] Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. 'Predicting Poverty and Wealth from Mobile Phone Metadata.' *Science* 350(6264): 1073–1076. <https://doi.org/10.1126/science.aac4420>
- [71] Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. 'The Parable of Google Flu: Traps in Big Data Analysis.' *Science* 343(6176): 1203–1205. <https://doi.org/10.1126/science.1248506>
- [72] Schelling, Thomas C. 1971. 'Dynamic Models of Segregation.' *Journal of Mathematical Sociology* 1(2): 143–186. <https://doi.org/10.1080/0022250X.1971.9989794>
- [73] Epstein, Joshua M., and Robert Axtell. 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press and MIT Press. <https://www.brookings.edu/books/growing-artificial-societies/>
- [74] Cook, R. Dennis. 1977. 'Detection of Influential Observation in Linear Regression.' *Technometrics* 19(1): 15–18. <https://doi.org/10.1080/00401706.1977.10489493>
- [75] Rubin, Donald B. 1976. 'Inference and Missing Data.' *Biometrika* 63(3): 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- [76] Silberzahn, Raphael, 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(3): 337–356. <https://doi.org/10.1177/2515245917747646>
- [77] Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.' *Psychological Science* 22(11): 1359–1366. <https://doi.org/10.1177/0956797611417632>
- [78] Open Science Collaboration. 2015. 'Estimating the Reproducibility of Psychological Science.' *Science* 349(6251): aac4716. <https://doi.org/10.1126/science.aac4716>
- [79] Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. 'The Preregistration Revolution.' *Proceedings of the National Academy of Sciences* 115(11): 2600–2606. <https://doi.org/10.1073/pnas.1708274114>

- [80] National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: National Academies Press.  
<https://doi.org/10.17226/25303>
- [81] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. ‘Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.’ *Science* 366(6464): 447–453.  
<https://doi.org/10.1126/science.aax2342>
- [82] Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. ‘Out of One, Many: Using Language Models to Simulate Human Samples.’ *Political Analysis* 31(3): 337–351.  
<https://doi.org/10.1017/pan.2023.2>
- [83] Bail, Christopher A. 2024. ‘Can Generative AI Improve Social Science?’ *Proceedings of the National Academy of Sciences* 121(21): e2314021121.  
<https://doi.org/10.1073/pnas.2314021121>
- [84] Squire, Peverill. 1988. ‘Why the 1936 Literary Digest Poll Failed.’ *Public Opinion Quarterly* 52(1): 125–133.  
<https://doi.org/10.1086/269085>
- [85] Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. ‘Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff.’ *Cambridge Journal of Economics* 38(2): 257–279. Published online December 24, 2013. Earlier version: PERI Working Paper 322, April 2013.  
<https://doi.org/10.1093/cje/bet075>
- [86] Fetters, Michael D., Leslie A. Curry, and John W. Creswell. 2013. ‘Achieving Integration in Mixed Methods Designs—Principles and Practices.’ *Health Services Research* 48(6, part 2): 2134–2156.  
<https://doi.org/10.1111/1475-6773.12117>
- [87] Vostokov, Dmitry. 2024. *Theoretical Software Diagnostics: Collected Articles*. Fourth Edition. OpenTask. ISBN 978-1-912636-91-4. Revision 4.01.
- [88] Vostokov, Dmitry. 2023. *Trace, Log, Text, Narrative, Data: An Analysis Pattern Reference for Information Mining, Diagnostics, Anomaly Detection*. Fifth Edition. OpenTask. ISBN 978-1-912636-58-7. Author-supplied Revision 5.06, August 2026.

