Part V

Applications

Chapter 23

Evaluating Science and Technology Research Institutes

This chapter employs Evaluatology to evaluate scientific and technological research institutes. I conceived the core concept, which Dr. Fanda Fan and I jointly implemented.

23.1 Introduction

As an independent entity or being affiliated with a university or company, a science and technology research institute (in short, STRI) plays an essential role as a driving force behind scientific and technological (S&T) progress. Past evaluation efforts regarding STRIs have been overly simplistic, primarily reducing their performance to mere quantification of publications, citations, or other bibliometric indicators. However, it only captures a narrow slice of the overall influence generated by scientific research institutions.

Within the discipline of *Evaluatology*, an STRI is formally treated as an *EO*. The essence of evaluation is to *uncover the effects* of an EO on a set of *AOs* under a clearly defined *SES*. From this perspective, the effects of an STRI should not be limited to academic influence alone. Instead, they encompass a broader spectrum of outcomes, including national development, human progress, and industrial advancement.

Moreover, the observable outcomes on these AOs—whether obtained through measurement or testing—are inevitably shaped by both the EOs and the *EXOs* within the SES. In practice, an AO such as a country's strategic alignment, an industry's innovation vitality, or humanity's sustainable development index reflects not only the direct effect transmitted from the EO, but also the derived or confounding influences introduced by EXOs, including talent resources, international cooperation, intellectual property protection, and the technology–capital environment. Evaluatology, therefore, requires that every measured or tested effect on an AO be decomposed along the EO $\rightarrow$ AO and EXO $\rightarrow$ AO pathways, ensuring that the outcome accurately isolates the effect attributable to the EO itself. Only through such precise attribution can the SES produce valid and unbiased judgments of an STRI's true effect on its AOs.

Finally, it is essential to recognize that the EO itself is not monolithic. An STRI consists of multiple internal components—such as talent development, evaluation mechanisms, public platforms, academic cooperation, management strategies, and technological specialization—each functioning as an essential component that generates its own effect pathways. These internal components jointly induce heterogeneous effects on the AOs, forming layered causal chains within the SES. A rigorous evaluation must therefore disentangle the contributions of each component of the EO, identify the internal effect mechanisms through which they influence different AOs, and determine how their interactions amplify or attenuate the overall effect of STRI. An SES for evaluating an STRI is shown in Figure 23.1.

The remainder of this chapter is organized to progressively deepen this perspective. Section 23.2 first reviews the academic-centric approaches that historically dominated the evaluation of an STRI, highlighting their limitations within the SES. Section 23.3 then expands the AO from narrow scholarly outputs to the broader impacts on national development, human progress, and industrial advancement. Section 23.4 develops the causal logic for isolating the genuine EO-induced effect from the influences of the EXOs. Finally, Section 23.5 decomposes the EO into its internal components and examines how these internal effect mechanisms co-produce the observable outcomes on the AOs. Together, these sections establish the causal and structural foundations for evaluating an STRI.

23.2 Traditional Evaluation Methodologies of Academic Achievements

For a long period, the evaluation of the STRI has been dominated by academic-oriented evaluation systems. Traditional frameworks equate S&T almost entirely with academic achievements, relying primarily on bibliometric indicators such as publication counts, citation impact, journal rankings, highly cited papers, and awards within the scientific community.

Core Journal Evaluation: The quality and influence of academic journals are commonly measured by a series of quantitative indicators, which are defined and published by various data providers.

- *Impact Factor (IF):* As the most established and widely recognized metric, the Impact Factor is published annually by Clarivate in its *Journal Citation Reports* (JCR). Its formula is:

$$\mathrm{IF}_{\text{year Y}} = \frac{\text{Citations in year Y to articles published in (Y-1) and (Y-2)}}{\text{Total citable items published in (Y-1) and (Y-2)}}. \quad (23.1)$$

  The Impact Factor reflects the average number of citations received by a journal's articles in the two years following publication. It has long been considered the "gold

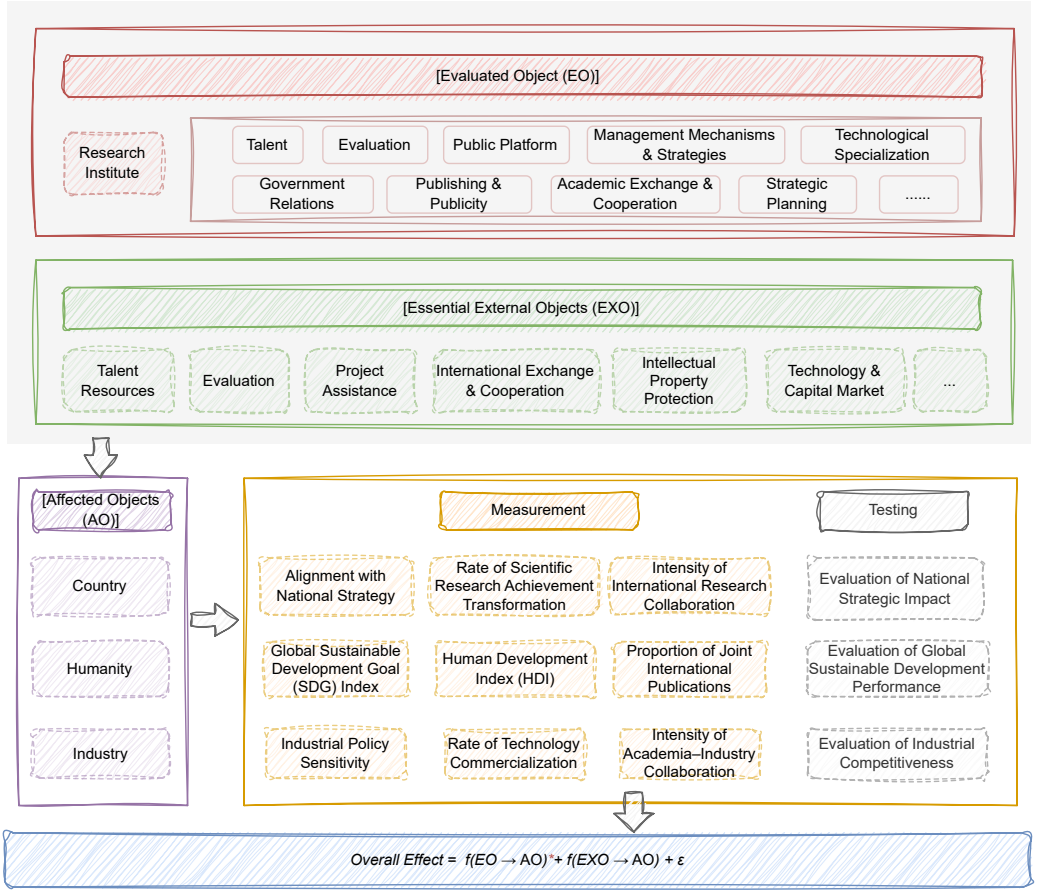

Figure 23.1: An SES for Evaluating an STRI.

standard" for journal quality, yet it faces criticism for its short calculation window, susceptibility to skew from a few highly cited articles, and lack of comparability across different scientific fields [66].

- *CiteScore:* Introduced by Elsevier based on its Scopus database, CiteScore is a major alternative to the IF. It utilizes a longer four-year window for both citations and publications and includes a broader range of document types (e.g., reviews, letters), aiming to provide a more comprehensive, transparent, and robust metric [199]. The formula is:

$$\text{CiteScore}_Y = \frac{\sum_{i=Y-4}^{Y-1} \text{Citations}_i}{\sum_{i=Y-4}^{Y-1} \text{Published Documents}_i}. \tag{23.2}$$

- *SCImago Journal Rank (SJR):* Also derived from the Scopus database, the SJR incorporates an algorithm similar to Google's PageRank [137]. It measures not

just the quantity but also the "quality" of citations, assigning a higher weight to citations from more prestigious journals. This allows SJR to measure the scientific prestige of a journal rather than just its raw citation traffic [57]. Due to its iterative nature, it does not have a simple fractional formula.

- *Source Normalized Impact per Paper (SNIP):* The SNIP metric is designed to address the challenge of cross-disciplinary comparisons. It normalizes a journal's raw citation impact by the "citation potential" of its specific subject field, thus measuring the relative impact of a paper within its domain. A SNIP value greater than 1.0 indicates that the journal's citation impact is higher than the average for its field [114]. The formula is expressed as:

$$\text{SNIP} = \frac{\text{Raw Impact per Paper (RIP)}}{\text{Relative Citation Potential (RCP)}}. \tag{23.3}$$

Journal Ranking and Partitioning: Beyond single metrics, journal partitioning provides a more intuitive hierarchical classification, helping researchers quickly evaluate a journal's standing within its discipline.

- *JCR Quartiles:* Published by Clarivate, this system ranks journals within a subject category based on their Impact Factor. The list is then divided into four equal parts: *Q1* (top 25%), *Q2* (25-50%), *Q3* (50-75%), and *Q4* (bottom 25%) [38].
- *CAS Partition:* The Chinese Academy of Sciences (CAS) partition is widely used in the Chinese academic community. It is based on a journal's three-year average IF and employs a "pyramid" distribution model. Within each discipline, the top 5% of journals are assigned to *Zone 1*, 6%-20% to *Zone 2*, 21%-50% to *Zone 3*, and the remainder to *Zone 4*. The most elite journals in Zones 1 and 2 are further designated as "*Top Journals*" [35].

Conference Ranking and Partitioning: In rapidly evolving fields such as Computer Science, top-tier academic conferences are often considered more prestigious than many journals due to their short review cycles and ability to disseminate cutting-edge research quickly. The evaluation of conferences typically relies on peer-based expert evaluation rather than a single quantitative formula.

- *CCF Recommended International Conference List:* Curated by the China Computer Federation (CCF), this list categorizes international conferences in computer science into three tiers: *A, B, and C*. Tier A represents the top-tier conferences with the highest academic impact. The evaluation criteria are multifaceted, considering a conference's history, review quality, paper acceptance rate, and overall influence. For example, the Conference on Neural Information Processing Systems (*NeurIPS*) is ranked as a *Tier A* conference [34].

- *CORE Ranking:* Published by the Computing Research and Education Association of Australasia, the CORE ranking is another internationally recognized system for computer science conferences. It classifies venues into four tiers: $A*$ (flagship), A (excellent), B (good), and C (standard). *NeurIPS* is ranked as $A*$ in this system [40].

In summary, the traditional evaluation of academic achievements heavy relies on a mature yet limited set of bibliometrics and ranking systems. Under this paradigm, the EO is implicitly judged through a narrowly defined subset of academic outputs, while the broader effects on national development, human progress, and industrial advancement remain largely unexamined. Consequently, these academic-centric methods capture only superficial manifestations of research activity rather than the full causal contributions of the research institute to its affected objects.

23.3 Beyond Academic Influence: An SES for STRIs

The SES constitutes the core evaluation model of Evaluatology. It formalizes how an EO produces observable effects on a set of AOs under specific interrogation conditions, while accounting for the influence of EXOs. Within this framework, STRI activities are interpreted through their effect pathways, allowing each measurable or testable outcome on an AO to be traced back to the EO, the EXOs, or their interactions. The SES therefore models the causal architecture of STRI by integrating three components—EO, EXO, and AO—into a unified structure that links institutional capability, environmental context, and societal effects with conceptual and methodological coherence.

At the top of the SES resides the EO, representing the entity under investigation that directly undertakes innovation and knowledge-creation activities. Each EO is composed of multiple internal components—*talent cultivation, research management, evaluation and incentive structures, public platforms, and technological specialization*—which together form its internal effect mechanisms. These components of EOs jointly shape the intrinsic capability of the organization to generate, transform, and disseminate scientific and technological knowledge, and they govern how the EO ultimately induces effects on its AOs within the SES.

Beyond the EO are the *EXOs*, which represent the objects that influence, constrain, or amplify the EO's ability to generate effects within the SES. Typical EXOs include *talent resources, funding mechanisms and policies, international cooperation, government relations, intellectual property protection, and technology–capital markets.* The EO–EXO interface delineates the dynamic boundary through which policy incentives, resource flows, and knowledge exchange operate, thereby shaping how the EO's internal capabilities are converted into measurable and testable effects on the AOs.

These effects ultimately materialize in the *AOs*, which constitute the domains that receive and exhibit the consequences of scientific and technological activity. Importantly, AOs should not be limited to academic outputs such as publications or citations. In the context of STRI, AOs encompass three higher-level spheres of societal influence:

- *National Development* —national competitiveness, strategic security, and policy alignment;
- *Human Progress* —knowledge, social welfare, sustainability, equity, and global well-being;
- *Industrial Advancement* —technological upgrading, productivity enhancement, and economic transformation.

Together, these domains reflect the full spectrum of effects that S&T can induce across civilization. A comprehensive evaluation of S&T must therefore quantify not only scholarly achievements but also the multi-domain consequences produced through the EO → AO, EXO → AO causal pathways within the SES.

To operationalize the SES for an STRI within the framework of Evaluatology, the evaluator relies on two fundamental interrogations: *measurement* and *testing*. Together, they attribute values to the observable effects on the AOs and verify whether propositions or models about the EO-induced effects conform to test oracles.

Measurement attributes values to the observable effects produced along the EO → AO and EXO → AO pathways under specified ECs. Each indicator corresponds to a measurable manifestation of these causal relationships. For example, the *rate of scientific research achievement transformation* and the *rate of technology commercialization* quantify how S&T generated by the EO propagate into industrial and economic AOs. Indicators such as the *intensity of international research collaboration* and the *proportion of joint international publications* capture cross-border knowledge flows and reflect the EO's contribution to global scientific exchange. Policy-oriented indicators—including *industrial policy sensitivity* and *alignment with national strategies*—measure how EO activities influence country-level AOs, while the *Human Development Index (HDI)* extends measurement to humanitarian AOs by linking scientific and technological progress to improvements in human well-being. Measurement thus relies on observable data from administrative records, research output databases, collaboration networks, policy documents, and socio-economic statistics, and converts them into comparable numerical quantities.

Testing is a verification process of running test cases to determine whether a proposition or a model about the EO's effect on its AOs conforms to a *test oracle*. In the context of STRI, a test oracle specifies the mandated or expected outcomes of S&T under a given SES—for example, a target level of *national strategic impact*, a benchmark for *global sustainable development performance*, or a required threshold of *industrial competitiveness*. A test case is a predefined interrogation condition, consisting of selected EO components, EXOs, AOs, time windows, and data samples, under which the measured indicators are computed. Testing executes these test cases and compares the actual measured outcomes with those mandated by the corresponding test oracles, yielding pass/fail decisions or acceptance/rejection of propositions such as "STRI is aligned with national strategies" or "STRI significantly promotes global sustainable development."

Through iterative cycles in which measurement provides quantitative inputs and testing verifies explicitly defined test oracles, the methodology ensures that evaluation

outcomes are both numerically grounded and logically consistent.

23.4 Accurate Attribution: Identifying the True Effect of EO

Within the SES, the causal structure of evaluation is inherently interconnected. Objects —including STRIs, journals, conferences, public platforms, and temporal environments —continuously interact and generate overlapping effects across different objects of the system. In this SES, the EXOs—such as policy incentives, funding programs, collaboration opportunities, and platform visibility—are not static backgrounds but active objects that induce their own effects on the AOs and modulate the effects originating from the EO. Consequently, any observable outcome on an AO represents a composite effect that includes true EO-induced influence, EXO-induced influence, and their interaction-driven derived effects.

Accurate attribution seeks to *isolate the true EO-induced effect* by disentangling these interwoven causal sources. Empirically, only the AO outcomes are directly observable. Let $\widehat{Y}_{\text{AO}}$ denote the measured effect on an AO, which aggregates contributions from multiple pathways:

$$\widehat{Y}_{\text{AO}} = f_{\text{EO}}(\text{EO} \to \text{AO}) + f_{\text{EXO}}(\text{EXO} \to \text{AO}) + f_{\text{int}}(\text{EO}, \text{EXO} \to \text{AO}) + \varepsilon, \tag{23.4}$$

where f_{EO} represents the true effectof the EO on the AO, f_{EXO} represents the effect induced by the EXOs, f_{int} captures the interface through which EXOs amplify, attenuate, or reshape the EO's effect on the AO, and ε represents the noise term. Because the evaluator can observe only $\widehat{Y}_{\text{AO}}$, identifying the true EO-induced effect requires a process of *causal reconstruction* [162] under explicitly defined interrogation conditions.

To separate the EO's contribution from that of the EXOs, the true effect of the EO can be expressed as:

$$\begin{aligned} \text{True Effect of the EO} = \mathbb{E}\Big[\widehat{Y}_{\text{AO}} \mid \text{EO} = 1, \text{EXO} = \text{constant}\Big] \\ - \mathbb{E}\Big[\widehat{Y}_{\text{AO}} \mid \text{EO} = 0, \text{EXO} = \text{constant}\Big], \end{aligned} \tag{23.5}$$

which conditions on a fixed EXO configuration. Conceptually, this corresponds to a counterfactual comparison under the same interrogation conditions: *How would the AO outcome appear if the EO's effect were absent?*

A variety of methodological approaches can be employed to perform this causal reconstruction under fixed interrogation conditions. One class of approaches relies on *effect decomposition* [3] methods, which partition the measured AO outcome into components attributable to the EO, the EXOs, and their interaction-induced derived effects. Another class of approaches adopts *controlled comparison strategies* [72], in which EO and non-EO objects are compared under identical EXO conditions to eliminate contextual variability. Additionally, *structural reconstruction techniques* [163]—such as constraint-based or score-based reconstruction of effect pathways—can be applied to infer the structure

of the EO → AO relationship from observational data. Finally, testing provides a way to verify the inferred effect. These approaches collectively ensure that the inferred effect of the EO reflects the true effect of the EO rather than the advantages or perturbations introduced by the surrounding EXOs.

In the framework of *Evaluatology*, accurate attribution elevates evaluation from descriptive comparison to a form of causal accountability. Rather than simply contrasting observed performances across STRIs, the evaluator examines why such performance arises by tracing effect pathways and identifying how EXOs modulate, enhance, or confound the EO-induced effects. For example, an STRI's high publication volume or strong technology-transfer performance may reflect genuine internal capability, or may instead be driven by favorable EXOs such as abundant funding, advantageous partnerships, or unique temporal conditions. Without isolating these sources, evaluations risk conflating contextual advantages with the intrinsic capability of the EO.

Ultimately, accurate attribution reframes S&T progress of STRIs as a context-adjusted causal effect, revealing the true effect generated by the EO within a shared evaluation environment. This principle provides the foundation for the next step: *tracing the internal components of the EO* to determine how its internal effect mechanisms collectively produce the measurable effects observed on AOs.

23.5 Tracing Internal Mechanisms: Component-Level Attribution within EO

After isolating the true effect of the EO from that of the EXOs, a further analytical step is required to understand *how* this effect is internally generated within the EO. An EO is not a monolithic object, but a structured system composed of multiple interdependent *internal components* that jointly determine its S&T progress. These components—such as *talent cultivation*, *evaluation and incentive structures*, *public service platforms*, *management mechanisms and strategies*, and *technological specialization*—function as internal effect mechanisms. Their coordinated interactions ultimately shape the EO's measurable and testable effects on the AOs within the SES.

Let $\mathbf{c}_{\mathrm{EO}} = \{c_1, c_2, \ldots, c_n\}$ denote the set of internal components of the EO. Each component c_i contributes both individually and jointly to the observed AO outcome $\widehat{Y}_{\mathrm{AO}}$. The overall EO-induced effect can therefore be expressed as:

$$f_{\mathrm{EO}}(\mathbf{c}_{\mathrm{EO}} \to \mathrm{AO}) = \sum_i g_i(c_i \to \mathrm{AO}) + \sum_{i<j} g_{ij}(c_i, c_j \to \mathrm{AO}) + \varepsilon_{\mathrm{EO}}, \tag{23.6}$$

where f_{EO} represents the true effect of the EO on the AO, g_i denotes the direct effect of component c_i, g_{ij} represents higher-order interaction-induced effects among components, and $\varepsilon_{\mathrm{EO}}$ captures residual internal effects that are not directly observable.

To evaluate the marginal contribution of each component under a given EXO con-

figuration, we consider the sensitivity of the AO outcome with respect to c_i:

$$\frac{\partial \widehat{Y}_{\mathrm{AO}}}{\partial c_i} = \underbrace{\frac{\partial f_{\mathrm{EO}}}{\partial c_i}}_{\text{direct EO component effect}} + \underbrace{\sum_k \frac{\partial f_{\mathrm{int}}}{\partial c_i} \frac{\partial x_k}{\partial c_i}}_{\text{EXO-mediated modulation}} . \tag{23.7}$$

Let $\widehat{Y}_{\mathrm{AO}}$ denote the measured effect on an AO. The first term represents the true contribution of the internal component, while the second term captures how EXOs modulate the component's effect—such as how funding levels, collaboration opportunities, or policy incentives amplify or attenuate the contribution of a particular mechanism. This differential formulation provides a quantitative basis for *component-level attribution*, clarifying how each internal mechanism shapes the overall EO-induced effect on the AOs.

From the perspective of *Evaluatology*, this analysis constitutes a form of mechanistic attribution. It shifts the evaluative question from "*How much true effect does this EO generate?*" to "*Which internal mechanisms generate the effect, and through what interactions?*" Such insight enables targeted STRI improvement, evidence-based policy design, and fairer cross-EO comparisons under heterogeneous EXOs.

At last, component-level attribution reveals that the S&T of an EO emerges not from isolated functions, but from the synergistic coordination of multiple internal mechanisms —each leaving a measurable causal footprint in the SES and collectively determining the EO's observable impact on the AOs.

23.6 Summary

This chapter presented that evaluating an STRI requires a shift from a bibliometric approach toward a causally grounded revealing of how an STRI or its components induce true effects within an STRI.

Chapter 24

Testbed Principles, Methodologies and Case Studies

This chapter formalizes what a testbed is and presents principles, methodology, and a case study of a testbed. I conceived the core concept, which Dr. Wanling Gao and I jointly implemented.

24.1 What is a Testbed?

Testbeds —whether conceived as experimental platforms, emulated environments, or full-fledged simulation systems —are indispensable tools for evaluating design choices and implementation trade-offs across engineering domains.

However, testbeds are not formally defined. I define the testbed as *an evaluation model that is designed and implemented for a class or different classes of cause objects or EO to simulate a perfect or imperfect, or simple SES, under which the effect of EOs could be accurately attributed.*

24.2 Testbed Principles

The essential purpose of a testbed is to enable controlled, repeatable, and interpretable experiments through which the causal effects of *EOs* on their corresponding *AOs* can be observed and quantified.

An ideal testbed is to simulate a perfect *SES* under which we can measure or test the effects of EO on AOs under different *EXOs*. According to the discussions in Section 14.2.1, a perfect SES has four unique characteristics: it can correctly recognize AOs and EXOs; it can completely isolate irrelevant objects; under a perfect SES, we can infer the true effect of the EO; we can freely manipulate the full space of SES.

Unfortunately, due to different limitations, we can only achieve imperfect SES in most cases. So, above all, a testbed should embody the principle of *controlled realism*: the ability to replicate the functional and causal relationships of an SES while providing researchers with sufficient control and observability to infer EO effects accurately.

Three guiding principles underlie the design of any testbed:

(1) Representativeness: A testbed must approximate a prefect or imperfect SES with sufficient fidelity such that results derived from it remain valid and generalizable to the actual system. Representativeness ensures that the essential relationships among EO, AO, and EXO are faithfully maintained, even if simplified. For example, a hardware simulator that preserves timing and dependency characteristics can yield representative insights even without physical circuitry.

(2) Controllability: A testbed should allow for explicit manipulation of both EO, AO, and EXO configurations while holding irrelevant variables constant. This capacity for controlled experimentation is what transforms an imperfect SES into a more analyzable model. In the ideal scenario (perfect SES), all irrelevant influences can be eliminated; in practice, the testbed approximates this condition as closely as feasible.

(3) Transparency and Repeatability: A testbed must support full visibility into its internal states and permit experiments to be replicated with deterministic or statistically bounded outcomes. Transparency ensures interpretability—researchers can trace observed results back to underlying causes—while repeatability ensures that results can be validated independently.

In essence, the testbed operationalizes Evaluatology's central aim: constructing a measurable, manipulable, and inferable environment that enables the transition from observation to causal understanding. Whether for a perfect, imperfect, or simple SES, every testbed serves as a concrete realization within the constraints of technology, knowledge, and resources.

24.3 Fundamental Testbed Methodologies

Building upon the principles above, testbed methodology defines how evaluators construct, operate, and refine testbeds to achieve reliable causal inference within practical constraints. The methodological foundation of testbeds rests upon their correspondence to the three SES types—perfect, imperfect, and simple—each representing a different trade-off between fidelity and feasibility.

(1) A Testbed Simulating a perfect SES (in short, *perfect testbed*): A perfect testbed represents the theoretical ideal scenarios where all irrelevant objects are fully isolated, and all relevant interactions are explicitly modeled. In such environments, researchers can infer the true EO effect because the causal structure is entirely transparent. For instance, in algorithmic benchmarking under a fully deterministic simulation, every input, random seed, and computational state could be changed and fixed, enabling perfect reproducibility. However, such testbeds are often unattainable in reality due to their prohibitive complexity and abstraction cost.

(2) A Testbed Simulating an imperfect SES (in short, *imperfect testbed*): An imperfect testbed approximates the real-world system but inevitably includes certain external factors that cannot be fully controlled or isolated. In other words, while the testbed seeks to capture the causal relationship between the EO and the AO, some influences from the surrounding environment or unobserved variables may remain. Although this lack of complete isolation introduces uncertainty into causal inference, it enables the evaluation to reflect more realistic and operational conditions.

For example, when evaluating CPUs under different environmental temperatures, the performance results may vary due to thermal effects. Such incomplete control—referred to as imperfect isolation—means that the influence of temperature cannot be entirely excluded. However, this variability also makes the results more representative of real-world usage. Hence, an imperfect testbed provides a pragmatic balance between causal rigor and ecological validity.

(3) A Testbed Simulating a simple SES (in short, *simple testbed*): Recognizing the infeasibility of exhaustive evaluation, a simple testbed reduces the complexity of the evaluation environment through both *sampling* and *simplification*.

Formally, a simple SES defines a reduced and sampled perfect or imperfect SES (detailed formalization in Section 14.2.3) that captures representative configurations. This subspace may be obtained through experimental design principles—such as factorial sampling, stratified selection, or Latin hypercube methods—to ensure diversity and coverage while controlling evaluation cost.

Beyond sampling, simplification can be achieved by abstracting or aggregating variables within the SES. For instance, rather than modeling every environmental parameter in detail, closely related variables (e.g., temperature and humidity) can be combined into a single composite factor; or, less influential EXOs can be fixed to typical values to focus on primary sources of variability. Such simplifications maintain the essential causal structure while reducing computational and experimental burden.

In essence, a simple SES is an *EO equipped with a simplified and sampled EC*. It trades completeness for tractability—omitting minor or redundant conditions—yet remains grounded in statistical validity and causal interpretability. By doing so, it enables efficient, scalable, and interpretable evaluation without losing sight of the underlying causal mechanisms.

(4) Evaluation Procedure: Across all SES types, the general procedure of testbed design and implementation consists of four canonical phases:

1. *Model Construction:* Define EO, AO, and EXO, and formalize their relationships within the testbed architecture.

2. *Condition Sampling:* Generate a representative EC set of C' that spans key variations in EXO and AO parameters.

3. *Outcome Measurement:* Execute controlled experiments to obtain outcome distributions $oe(o|c')$(detailed formalization in Section 12.1.1), accounting for stochastic variability through repetition.

4. *Effect Inference:* Apply statistical analysis (e.g., ANOVA, regression, or covariance decomposition) to estimate the inferred effects of the EO on AOs.

5. *Hypothesis Testing:* Perform a hypothesis test on the inferred effects of the EO on AOs.

This structured methodology provides a unifying framework: perfect testbeds guarantee theoretical validity; imperfect testbeds offer empirical realism; and simple testbeds ensure scalability. Together, they form a methodological continuum that adapts Evaluatology to both scientific inquiry and engineering application.

24.4 Case Studies

To illustrate the application of testbed principles in practice, we examine representative cases across distinct evaluation domains, demonstrating how different SES types and testbed methodologies are instantiated.

Case 1: CPU Performance Evaluation: In hardware performance benchmarking, the *EO* is the CPU, the *AO* is the computing system (including OS, memory, and disk), and the *EXO* consists of workloads, datasets, and compilers.

Different testbed exemplifies Ealuatology's balance between rigor and feasibility.

A perfect testbed provides the means to evaluate a CPU while isolating and exploring all AO and EXO space—an unattainable ideal in practice.

An imperfect testbed provides the means to evaluate a CPU by executing standardized benchmarks (e.g., SPEC CPU [181]) under controlled but not fully isolated conditions on limited AOs.

A simple testbed, such as cloud-based benchmarking platforms, samples representative workloads across configurations on a fixed AO and applies statistical inference to estimate CPU-specific performance while accounting for environmental noise.

Case 2: Drug Efficacy Evaluation: In biomedical evaluation, the *EO* is the drug compound, the *AO* is the human body, and the *EXO* includes diet, stress, and environmental exposure. A perfect testbed corresponds to a theoretical physiological model with a fully controllable AO and EXO that completely isolates the drug's biochemical effects—impossible in reality.

Clinical trials thus represent imperfect testbeds, where randomization and blinding serve as tools to approximate equivalent evaluation conditions.

A simple SES arises in simulation-based pharmacokinetics, where population-based sampling models the drug's effect across synthetic patient cohorts, providing scalable yet interpretable estimates of efficacy.

Discussion: Across domains, these case studies reveal a recurring trade-off: the cost increases with fidelity. Thus, the practical art of testbed design lies in constructing *simple SESs*—testbeds that retain essential causal structures while remaining operationally feasible. Such testbeds operationalize Evaluatology's fundamental vision: transforming abstract causal reasoning into reproducible, evidence-based evaluation that bridges the gap between theory and practice.

24.5 Summary

This chapter establishes a unified theoretical and methodological foundation for testbeds within the framework of Evaluatology.

Chapter 25

Evaluatology-based Artificial Intelligence

In this chapter, we begin by defining the fundamental concepts, assumptions, and problem formulations that ground artificial intelligence (AI). Among various AI paradigms, we focus on the prevailing data-driven deep learning paradigm. Despite its empirical success, this paradigm remains a black box: it can judge whether outcomes are good or bad but provides little understanding of why they occur or how models can be systematically improved.

I conceived the core concept, which Dr. Guoxin Kang, Dr. Wanling Gao, and I jointly implemented.

25.1 The Limitations of Existing AI Paradigms

Early AI was dominated by the symbolic paradigm, grounded in the belief that intelligence could be fully captured through symbolic logic and explicit rules [131, 130, 133, 67]. This paradigm laid the conceptual foundation for the Turing Test [205], defining intelligence as the capacity for symbolic manipulation. Expert systems represented the practical culmination of symbolic AI, encoding human knowledge into rule-based engines [27, 46, 117, 42]; however, they suffered from limited scalability and an inability to learn from data.

These limitations catalyzed the rise of the connectionist paradigm, which is inspired by biological neural networks [161, 81, 84]. Hebb's seminal theory linked synaptic adaptation to learning, providing a theoretical bridge between neuroscience and machine learning [79]. Building on multilayer neural architectures and efficient training algorithms, deep learning emerged by enabling models to automatically discover patterns and statistical regularities from large-scale data [106, 76, 110, 210, 49].

In contrast to the symbolic paradigm and early connectionist advances, modern AI has been shaped by a data-driven deep learning paradigm, which assumes that intelligence can be approximated by learning statistical regularities from massive datasets [98]. This data-centric principle has reached its most visible success in large language models

(LLMs) [25], whose performance scales predictably with training data volume, model size, and compute budget [99]. However, this scaling paradigm is increasingly constrained by a looming data bottleneck. As high-quality human-authored data becomes saturated and expensive to curate, synthetic data generation has emerged as a promising alternative.

Despite its scalability, synthetic data introduces a new layer of complexity [179, 125, 14]. Crucially, the quality of synthetic data is fundamentally limited by the generative models that produce it, which are often black-box architectures with little transparency or interpretability. This lack of visibility makes it difficult to trace the root causes of errors or biases in downstream models back to specific properties of the synthetic data. When performance deteriorates, it remains unclear whether the issue lies in data coverage, semantic consistency, or deeper representational flaws.

In practice, current synthetic data suffers from several well-documented issues: 1) generative models may fail to match the statistical distribution of real data, introducing biases that impair generalization. 2) Synthetic samples often contain logical contradictions or distorted features that are difficult to detect but can corrupt pre-training. Low diversity and mode collapse: generators tend to produce samples with limited variation, leading to models that overfit narrow modes and underperform on real-world variability.

To improve the quality, reliability, and usefulness of synthetic data, it is imperative to enhance the interpretability and evaluation of generative models. Without understanding what a generator has learned, and what it systematically omits, scaling synthetic corpora becomes a blind process, susceptible to spurious correlations and misalignment.

These observations motivate a shift toward an Evaluatology-based AI paradigm, in which systematic attribution and interpretability are not afterthoughts but central components of the AI development cycle. Regardless of the data source, all data are inherently generated under specific conditions. However, prevailing AI training methods largely ignore these generative conditions and focus exclusively on the data themselves. Such a deficiency leads to uneven and difficult-to-evaluate data quality, constrains interpretability and the capacity for causal discovery, and renders models fragile in the face of novel scenarios.

Our research intuition is that explicitly incorporating both data and their generative conditions into the training process can substantially enhance the effectiveness and transparency of AI. Even under limited data availability, leveraging the interplay between data and conditions allows the discovery of deeper causal structures, enabling models to capture the invariant informational essence beneath data diversity. By grounding learning in condition-aware causal relationships, we move toward more interpretable, attributable, and genuinely intelligent systems.

25.2 Basic Concepts and Principles of Deep Learning

We introduce the foundational concepts and principles of the data-driven deep learning paradigm.

25.2.1 Basic Concepts

Model Architecture. In the data-driven paradigm, the model architecture typically refers to deep neural networks, which serve as function approximators mapping inputs to outputs. These architectures are designed to scale with data volume and computational resources [111, 210, 71].

Dataset. A dataset comprises a large collection of labeled or unlabeled samples, used to train the model [106, 77, 197].

Loss Function. A loss function $\mathcal{L}$ quantifies the prediction error between the model output and the ground truth. Training aims to minimize this error over the dataset, i.e., $\min_\theta \mathcal{L}(\mathcal{D};\theta)$, enabling the model to learn the input–output mapping [161, 71, 36, 18].

25.2.2 Basic Principle

Given sufficiently large training data $\mathcal{D}$, sufficiently large model capacity (i.e., number of parameters) θ, and sufficient compute budget $\mathcal{C}$, a deep learning model f_θ is assumed to be capable of solving increasingly complex real-world tasks $\mathcal{T}$ via empirical risk minimization [208, 110, 71, 99, 83]:

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\mathcal{D};\theta), \tag{25.1}$$

where $\mathcal{L}$ is a loss function and $\hat{\theta}$ denotes the parameters of the optimized model obtained by minimizing the empirical loss. The model performance is typically evaluated by aggregate statistical metrics such as accuracy:

$$\text{Performance} = \mathcal{M}(f_{\hat{\theta}}, \mathcal{T}), \tag{25.2}$$

where $\mathcal{M}$ denotes a statistical measurement (i.e., a quantitative performance metric) and the compute budget $\mathcal{C}$ is assumed to scale proportionally with the model parameters and training data [99]:

$$\mathcal{C} \propto \theta \cdot \mathcal{D}. \tag{25.3}$$

However, these metrics are often treated as black-box indicators and offer limited causal interpretability [116, 52, 141].

25.3 The New AI Paradigm Based on Evaluatology

The Evaluatology-based AI paradigm constructs an SES, shifting AI research from merely answering *"Is the model good?"* to systematically address *"Under what ECs is it good?"*, *"Why is the model good?"*, and *"Which key design changes can make the model better?"*. As shown in Figure 25.1, this paradigm moves toward AI systems that are not only interpretable and causally attributable but also capable of more general intelligence. The following sections introduce the core elements of the Evaluatology-based AI paradigm and its four-step frameworks, which offer a promising path toward genuine general intelligence.

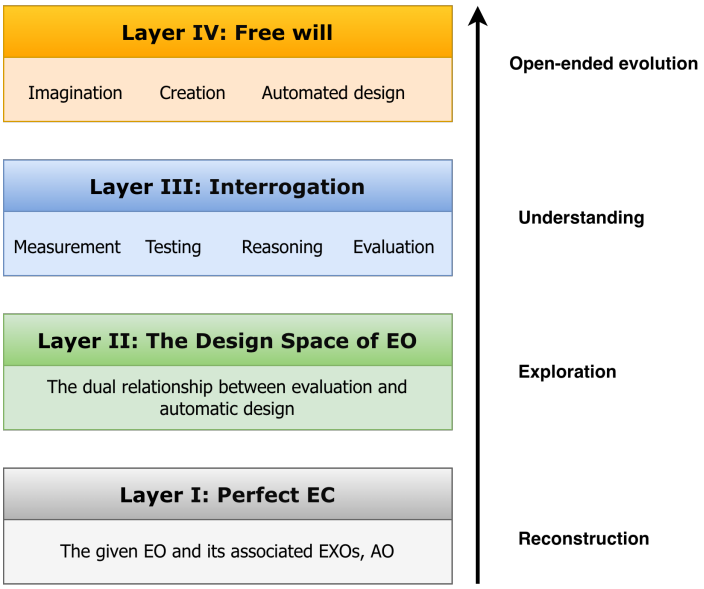

Figure 25.1: Evaluatology-based pathway toward strong AI.

25.3.1 Core Components of the SES

Section 12.1.2 formalizes the design problem in Evalutology. The *EO* is the object to be evaluated and designed, which manifests itself in various forms, including algorithmic structures such as a video retrieval network, an encoder-decoder architecture, or a recommendation algorithm, as well as core computer system components such as a CPU or a database system. The purpose of design is to search the specific *EO* configuration that achieves the optimal overall effect

The *EXOs*, together with the EO, jointly determine the overall effect on the AO. These include the training data, experimental configurations, hyperparameters, and environmental factors that define the context in which the model operates.

Please note that in Section 3, we defined data as "raw interrogation outcomes or their derived ones in different interrogation conditions." Every data sample, whether observational or experimental, must be generated under explicit interrogation conditions that specify the scene, data collection process, potential biases, and evaluation metrics used. This ensures causal traceability and reproducibility.

The *AO* reflects the measurable outcome or behavior influenced by both the EO and EXOs. It often corresponds to the computer system's measurable performance on downstream tasks, such as accuracy, latency, or robustness in deployment environments. The overall effect refers to the impact on the AO caused either by the design of the EO or by variations in the EXOs.

25.3.2 Structured Frameworks for Advancing to Strong AI

Building upon these established definitions, as shown in Figure 25.1, the Evaluatology-based paradigm instantiates them within the context of AI to develop intelligence through a progressive path, each reflecting a distinct relationship among the EO, EXOs, and AO.

Step I: Design and Implement a Perfect EC: At this foundational level, for any given EO and its associated EXOs and AO, a theoretically complete real-world distribution exists—that is, all possible ECs under which the training data could be generated. If sufficient resources such as time or computational power were available, this distribution could be exhaustively traversed, in principle. This step corresponds to *conditional brute-force computation*, which establishes the empirical foundation of intelligence by covering the entire EC space, although at high cost.

Step II: Explore the Design Space of EO Under a Perfect EC: Under a perfect EC, AI begins to explore the design space of an EO to identify potential design possibilities that faithfully reflect real-world behavior. The exploration typically proceeds in three steps: brute-force ensures exhaustive coverage of the design space, heuristic approaches leverage prior causal understanding and empirical knowledge to focus on high-potential regions, and pruning removes redundant or unproductive design paths to improve efficiency and convergence. Together, these steps enable AI to explore the design space systematically and efficiently at lower cost, preparing the ground for high complexity exploration under simple ECs.

Step III: Achieve High Complexity of Interrogation: After acquiring the ability to explore the design space, it advances into the stage of interrogation, engaging in epistemic inquiry through measurement, testing, reasoning, and evaluation. Guided by stakeholder requirements and under a fixed EO, AI systematically explores the EC space defined by the EXOs and AO to separate the effect of different objects and enable causal attribution. Through this process, it decomposes the overall effect on the AO into the respective effects of the EO and the EXOs, while refining the ECs to identify the ECs that most significantly influence performance.

Step IV: Achieve High Degree of Free Will: At this step, AI advances from causal understanding to intentional imagination, creation, and autonomous design. Supported by higher-order cognitive mechanisms such as counterfactual simulation, generative composition, and self-evaluation. Guided by the fixed simple ECs derived from the previous step, it first *imagines* alternative possibilities grounded in learned causal principles, then *creates* new designs of the EO through generative models, and ultimately performs *automatic design*—the process of finding a specific *EO* configuration to achieve optimal overall effect. Through this progression, AI demonstrates free and intentional decision-making, achieving creative generalization across contexts.

25.3.3 Summary of the Four Steps

The four steps outline a progressive path for the Evaluatology-based AI paradigm. The first step establishes the perfect EC. The second step explores the full design space of the EO under perfect EC. The third step interrogates under simple ECs to separate the effects of the EO and the EXOs. The fourth step enables intentional and autonomous design within fixed, simple ECs. Together, the four steps articulate a path for advancing AI toward an interpretable, self-improving, and causally grounded form of intelligence.

25.4 Case Study

To illustrate how the Evaluatology-based AI paradigm can advance database automatic design, we present the following case study.

In database automatic design [33], Evaluatology begins by defining the perfect EC. The *EO* is the database index, the *AO* corresponds to a minimally independent running database system, and the *EXOs* consists of all factors that influence index performance. The EXOs include, but are not limited to, data distribution and skew patterns, schema evolution and update frequencies, storage layout, and compression rules. Inspired by CPU Evaluatology, the EXOs are not fixed [213]; instead, workload and access distributions dynamically adapt to stakeholder requirements, reflecting realistic production variability rather than relying on a static benchmark.

Under the perfect EC, the full design space of the EO is explored. In a row-store database, accelerating access generally requires building indexes on selected columns. For a table with n columns, allowing arbitrary choices of column subsets and orders leads to a combinatorial design space of

$$\sum_{k=1}^{n} \frac{n!}{(n-k)!}, \tag{25.4}$$

which grows factorially and becomes computationally intractable. Here, $k = 1, 2, \ldots, n$ denotes the number of columns included in an index. AI examines this large space using a combination of brute-force enumeration to approximate completeness, heuristic exploration to focus on promising index patterns, and pruning to eliminate redundant or unproductive design paths.

From this exploration, the perfect EC is distilled into simple ECs that more faithfully simulate realistic deployment scenarios. These simple ECs capture factors such as mixed read/write ratios, skewed query distributions, and hardware-dependent cost models, allowing AI to analyze index performance under resource-aware and stakeholder-specific conditions.

With simple ECs established, AI performs interrogation through the four fundamental modes. Measurement quantifies index performance across workload variations; testing validates behavioral stability under simple ECs; evaluation integrates empirical evidence and reasoning to infer the true effect of each candidate index; and reasoning

explains why certain index structures lead to performance gains or regressions. This epistemic process produces a scientifically interpretable understanding of how index design factors shape system performance.

Finally, in the step of free will, the EO gains the capability for intentional redesign. Guided by the causal principles uncovered during interrogation, AI autonomously imagines alternative index forms, creates new structural variants, and performs automatic design to generate indexes that best satisfy stakeholder requirements. This moves beyond selecting from existing templates, such as B-tree, hash, or bitmap indexes, and enables the invention of novel index structures, achieving Evaluatology-driven database intelligence.

25.5 Summary

This chapter presented that the Evaluatology-based AI paradigm provides a new pathway beyond these constraints by redefining intelligence as a progressive path across a four-step evolution. This path establishes a dual relationship between evaluation (fixing EO, varying EC) and automatic design (fixing EC, varying EO), outlining a promising path for AI to evolve from an opaque data-driven black box toward an interpretable, causally grounded, and self-improving form of general intelligence.

Bibliography

[1] system. https://www.merriam-webster.com/dictionary/system. Accessed: February 6, 2024.

[2] Marvin C Alkin. Evaluation theory development. *Evaluation of short-term training in rehabilitation*, pages 9–16, 1970.

[3] Duane F Alwin and Robert M Hauser. The decomposition of effects in path analysis. *American sociological review*, pages 37–47, 1975.

[4] Bjørn Andersen, Tom Fagerhaug, Stine Randmæl, Jürgen Schuldmaier, and Johann Prenninger. Benchmarking supply chain management: finding best practices. *Journal of Business & Industrial Marketing*, 14(5/6):378–389, 1999.

[5] James C Anderson and David W Gerbing. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, 103(3):411, 1988.

[6] Aristotle. *The Organon*. Cambridge, Mass, London, 1938.

[7] Aristotle Aristotle, Aristotle, and CDC Reeve. *Metaphysics*, volume 1. Harvard University Press Cambridge, MA, 1933.

[8] Michaël Armand, Germain Faure, Benjamin Grégoire, Chantal Keller, Laurent Théry, and Benjamin Werner. A modular integration of sat/smt solvers to coq through proof witnesses. In *International Conference on Certified Programs and Proofs*, pages 135–150. Springer, 2011.

[9] Matthias Baaz and Christian G Fermüller. Resolution-based theorem proving for many-valued logics. *Journal of Symbolic Computation*, 19(4):353–391, 1995.

[10] Leo Bachmair and Harald Ganzinger. Resolution theorem proving. *Handbook of automated reasoning*, 1(02), 2001.

[11] Alexander Backlund. The definition of system. *Kybernetes*, 29(4):444–451, 2000.

[12] Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.

[13] Luciano Baresi and Michal Young. Test oracles. 2001.

[14] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.

[15] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958.

[16] Guillaume Bigourdan. Sur la mesure de la méridienne de france, à la fin du xviiie siècle, pour la détermination du mètre. *Bulletin astronomique, Observatoire de Paris*, 25(1):78–80, 1908.

[17] IEC BiPM, ILAC IFCC, IUPAP IUPAC, and OIML ISO. The international vocabulary of metrology—basic and general concepts and associated terms (vim). *JCGM*, 200:2012, 2012.

[18] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[19] Björn Blom and Stefan Morén. Analysis of generative mechanisms. *Journal of critical realism*, 10(1):60–79, 2011.

[20] David Bohm. *Quantum theory*. Courier Corporation, 2012.

[21] Kenneth A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

[22] Robert F Boruch and David S Cordray. An appraisal of educational program evaluations: Federal, state, and local agencies. 1980.

[23] Gregory Breit and John A Wheeler. Collision of two light quanta. *Physical Review*, 46(12):1087, 1934.

[24] Timothy A. Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2nd edition, 2015.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[26] James C Browne. An analysis of measurement procedures for computer systems. *ACM SIGMETRICS Performance Evaluation Review*, 4(1):29–32, 1975.

[27] Bruce G Buchanan and Edward A Feigenbaum. Dendral and meta-dendral: Their applications dimension. In *Readings in artificial intelligence*, pages 313–322. Elsevier, 1981.

[28] Robert C Camp. *Benchmarking: the search for industry best practices that lead to superior performance*. Asq Press, 1989.

[29] Donald T Campbell and HW Riecken. Quasi-experimental design. *International encyclopedia of the social sciences*, 5(3):259–263, 1968.

[30] Olivier Carnal and Jürgen Mlynek. Young' s double-slit experiment with atoms: A simple atom interferometer. *Physical review letters*, 66(21):2689, 1991.

[31] Henry Cavendish. Xxi. experiments to determine the density of the earth. *Philosophical Transactions of the Royal Society of London*, (88):469–526, 1798.

[32] David F Cavers. The food, drug, and cosmetic act of 1938: its legislative history and its substantive provisions. *Law & Contemp. Probs.*, 6:2, 1939.

[33] Sunil Chakkappen, Shreya Kunjibettu, Daniel McGreer, Masoomeh Javidi Kishi, Hong Su, Mohamed Ziauddin, Mohamed Zait, Zhan Li, and Yuying Zhang. Automatic indexing in oracle. *Proceedings of the VLDB Endowment*, 18(12):4924–4937, 2025.

[34] China Computer Federation. Ccf recommended international conference list, 2022.

[35] Chinese Academy of Sciences. Cas journal partition. http://www.fenqubiao.com, 2024.

[36] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

[37] Dinesh Choudhary and Vijay Kumar. Software testing. *Journal of Computational Simulation and Modeling*, 1(1):1, 2011.

[38] Clarivate Analytics. Journal citation reports. https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/, 2024.

[39] Evaluation Research Society Standards Committee et al. Evaluation research society standards for program evaluation. *Standards for evaluation practice. New directions for program evaluation*, (15):7–19, 1982.

[40] Computing Research and Education Association of Australasia. Core conference portal, 2024.

[41] Conférence Générale des Poids et Mesures. 11th general conference on weights and measures, 1960.

[42] Robert G Cowell, A Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic networks and expert systems*. Springer, 1999.

[43] David Roxbee Cox. Planning of experiments. 1958.

[44] Lee J Cronbach. Course improvement through evaluation. *Teachers college record*, 64(8):1–13, 1963.

[45] Lee J Cronbach, Sueann Robinson Ambron, Sanford M Dornbusch, Robert D Hess, Robert C Hornik, Denis Charles Phillips, Decker F Walker, and Stephen S Weiner. *Toward reform of program evaluation*. JSTOR, 1980.

[46] Randall Davis. Expert systems: Where are we? and where do we go from here? *AI magazine*, 3(2):3–3, 1982.

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.

[48] Bureau International des Poids et Mesures. *The International System of Units (SI) —9th edition*. Bureau International des Poids and Measures (BIPM), Sévres, France, 2019. Published May 2019; updated versions exist.

[49] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[50] Namiot Dmitry, Ilyushin Eugene, and Chizhov Ivan. On a formal verification of machine learning systems. *International Journal of Open Information Technologies*, 10(5):30–34, 2022.

[51] Jack J. Dongarra, Piotr Luszczek, and Antoine Petitet. The linpack benchmark: past, present and future. *Concurrency & Computation Practice & Experience*, 15(9):803–820, 2010.

[52] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[53] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

[54] Samuel Eilenberg and Saunders Mac Lane. General theory of natural equivalences. *Transactions of the American Mathematical Society*, 58:231–294, 1945.

[55] Elliot W Eisner. On the uses of educational connoisseurship and criticism for evaluating classroom life. *Teachers College Record*, 78(3):1–11, 1977.

[56] Junping Qiu et al. *Evaluation Science: Theory, Method and Practice*. Science Press, 1nd edition, 2010.

[57] Matthew E Falagas, Vasilios D Kouranos, Ricardo Arencibia-Jorge, and Drosos E Karageorgopoulos. Comparison of scimago journal rank indicator with journal impact factor. *The FASEB journal*, 22(8):2623–2628, 2008.

[58] Lawrence Fisher. Some new stock-market indexes. *The Journal of Business*, 39(1):191–225, 1966.

[59] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[60] Ronald Aylmer Fisher. *The design of experiments*. Springer, 1971.

[61] James Franck and Gustav Hertz. Über zusammenstöße zwischen elektronen und den molekülen des quecksilberdampfes und die ionisierungsspannung desselben. *Physikalische Blätter*, 23(7):294–301, 1967.

[62] Torkel Franzén. *Gödel's theorem: an incomplete guide to its use and abuse*. AK Peters/CRC Press, 2005.

[63] Gordon Fraser, Franz Wotawa, and Paul E Ammann. Testing with model checkers: a survey. *Software Testing, Verification and Reliability*, 19(3):215–261, 2009.

[64] Gottlob Frege. *Begriffsschrift: A Formula Language, Modeled upon that of Arithmetic, for Pure Thought*. Halle: Louis Nebert, 1879.

[65] Mike Furr. Scale construction and psychometrics for social and personality psychology. *Scale Construction and Psychometrics for Social and Personality Psychology*, pages 1–160, 2011.

[66] Eugene Garfield et al. The impact factor. *Current contents*, 25(20):3–7, 1994.

[67] Michael R Genesereth and Nils J Nilsson. *Logical foundations of artificial intelligence*. Morgan Kaufmann, 2012.

[68] Gene V Glass. *The growth of evaluation methodology*. Number 27. Laboratory of Educational Research, University of Colorado, 1969.

[69] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

[70] Weiwei Gong and Xu Zhou. A survey of sat solver. In *AIP Conference Proceedings*, volume 1836, page 020059. AIP Publishing LLC, 2017.

[71] Ian Goodfellow. Deep learning, 2016.

[72] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277, 2021.

[73] David J. Griffiths. *Introduction to Quantum Mechanics*. Pearson, 2 edition, 2005.

[74] Egon G Guba and Yvonna S Lincoln. *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. Jossey-Bass, 1981.

[75] Egon G Guba and Yvonna S Lincoln. *Fourth generation evaluation*. Sage, 1989.

[76] BI Guo-Qiang. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic type. *The Journal Neuroscience*, 18(24):10464–10472, 1988.

[77] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12, 2009.

[78] Daniel M Hausman and James Woodward. Independence, invariance and the causal markov condition. *The British journal for the philosophy of science*, 50(4):521–583, 1999.

[79] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.

[80] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[81] John A Hertz. *Introduction to the theory of neural computation*. Crc Press, 2018.

[82] David Hilbert and Wilhelm Ackermann. *Grundzüge der theoretischen Logik*. Springer, 1934.

[83] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[84] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[85] Ernest R House. *Evaluating With Validity. Beverly Hills*. California: Sage Publications, 1980.

[86] Colin Howson. Popper, prior probabilities, and inductive inference. *The British journal for the philosophy of science*, 38(2):207–224, 1987.

[87] Rick H Hoyle. *Structural equation modeling: Concepts, issues, and applications*. Sage, 1995.

[88] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, Berlin, 2005.

[89] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

[90] OT Inan, P Tenaerts, SA Prindiville, HR Reynolds, DS Dizon, K Cooper-Arnold, M Turakhia, MJ Pletcher, KL Preston, HM Krumholz, et al. Digitizing clinical trials. *NPJ digital medicine*, 3(1):101, 2020.

[91] Raj Jain. *The art of computer systems performance analysis*, volume 182. John Wiley & Sons Chichester, 1991.

[92] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[93] Jean Jenkins and Susan Hubbard. History of clinical trials. In *Seminars in oncology nursing*, volume 7, pages 228–234, 1991.

[94] Lizy Kurian John and Lieven Eeckhout. *Performance evaluation and benchmarking*. CRC Press, 2018.

[95] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, N.J., 6 edition, 2007.

[96] Karl G. Jöreskog. A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2):i–41, 1970.

[97] Raghu N Kacker. On quantity, value, unit, and other terms in the jcgm international vocabulary of metrology. *Measurement Science and Technology*, 32(12):125015, 2021.

[98] Guoxin Kang, Wanling Gao, and Jianfeng Zhan. Evaluatology-driven artificial intelligence. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, page 100245, 2025.

[99] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[100] Roger E Kirk. Experimental design. *Sage handbook of quantitative methods in psychology*, pages 23–45, 2009.

[101] Rex B Kline. *Principles and practice of structural equation modeling*. Guilford publications, 2023.

[102] Michael E Knudson. A performance measurement and system evaluation project plan proposal. *ACM SIGMETRICS Performance Evaluation Review*, 13(1):20–31, 1985.

[103] Andrey Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. Cited for the axiomatic definition of Probability, Random Variable, and Distribution Function.

[104] Samuel Kounev, Klaus-Dieter Lange, and Joakim Von Kistowski. *Systems Benchmarking*. Springer, 2020.

[105] Robert Kowalski. Logic programming. In *Handbook of the History of Logic*, volume 9, pages 523–569. Elsevier, 2014.

[106] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[107] Christoph Kröger. Evaluation: Definitions and concept. *1998). Evaluation drug prevention in the European union. Scientific monoghraph series*, (2):61–66, 1998.

[108] Sean P Lally. Henry cavendish and the density of the earth. *The Physics Teacher*, 37(1):34–37, 1999.

[109] Antoine Lavoisier. *Traité Élémentaire de Chimie*. Cuchet, Paris, 1789.

[110] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[111] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[112] Adam J Lee and Marianne Winslett. Enforcing safety and consistency constraints in policy-based authorization systems. *ACM Transactions on Information and System Security (TISSEC)*, 12(2):1–33, 2008.

[113] T. D. Lee and C. N. Yang. Question of parity conservation in weak interactions. *Phys. Rev.*, 106:1371–1371, Jun 1957.

[114] Loet Leydesdorff and Tobias Opthof. Scopus's source normalized impact per paper (snip) versus a journal impact factor based on fractional counting of citations. *Journal of the American society for information science and technology*, 61(11):2365–2369, 2010.

[115] Peter Lipton. Inference to the best explanation. *A Companion to the Philosophy of Science*, pages 184–193, 2017.

[116] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[117] Peter JF Lucas and Linda C Van Der Gaag. *Principles of expert systems*. Addison Wesley Longman, 1991.

[118] Sandra Mathison. *Encyclopedia of evaluation*. Sage publications, 2004.

[119] William Anderson McCall. *How to experiment in education*. Macmillan, 1926.

[120] Gregor Mendel. Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*, 1865.

[121] Tim Menzies. Applications of abduction: knowledge-level modelling. *International journal of human-computer studies*, 45(3):305–335, 1996.

[122] Albert Messiah. *Quantum mechanics*. Courier Corporation, 2014.

[123] David Moher, Kenneth F Schulz, Douglas G Altman, and CONSORT Group*. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of internal medicine*, 134(8):657–662, 2001.

[124] Sara Monti, Vittorio Grosso, Monica Todoerti, and Roberto Caporali. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology*, 57(Supplement_7):vii54–vii58, 2018.

[125] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in computer vision. *arXiv preprint arXiv:2403.10075*, 2024.

[126] Vikas Nagaraj. Automating test vector validation for silicon verification at scale. *International Journal of Engineering and Architecture (IJEA)*, 2(1):76–113, 2025.

[127] Leland Gerson Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

[128] D Nevo, A Lewy, S Kugelmass, G Ben-Shakar, N Blass, RF Boruch, DJ Davis, B Nevo, D Nevo, P Tamir, et al. The evaluation of a multi-dimensional project. *Lewy, A.(et al.) Decision Oriented Evaluation In Education, International Science Services*, 1981.

[129] David Nevo. The conceptualization of educational evaluation: An analytical review of the literature. *Review of Educational Research*, 53(1):117–128, Spring, 1983.

[130] Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, page 1959. Pittsburgh, PA, 1959.

[131] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

[132] Jerzy Neyman and Egon S. Pearson. The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29:492–510, 1933.

[133] Nils J Nilsson. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.

[134] Emmy Noether. Idealtheorie in Ringbereichen. *Mathematische Annalen*, 83:24–66, 1921.

[135] National Institute of Standards and Technology. Meter bar 27, n.d. Accessed: 2025-10-26.

[136] US General Accounting Office. Assessing social program impact evaluations: A checklist approach, 1978.

[137] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.

[138] Michael Quinn Patton. *Utilization-focused evaluation. Beverly Hills*. Ca: Sage, 1978.

[139] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition, 2000. Cited for the formal definition of Causality and the do-operator.

[140] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.

[141] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.

[142] Charles Sanders Peirce. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press, 1934.

[143] Rupert Pennick. The history of the metric system. *Journal of Geomancy*, 2(3):61–65, April 1978.

[144] Ahti-Veikko Pietarinen. Abduction and diagrams. *Logic Journal of the IGPL*, 29(4):447–468, 2021.

[145] M Provus. Evaluation as public policy. *Curriculum Theory Network*, 3(8-9):33–44, 1972.

[146] Sreeram V Ramagopalan, Alex Simpson, and Cormac Sammon. Can real-world data really replace randomised clinical trials? *BMC medicine*, 18(1):1–2, 2020.

[147] Ramakrishnan Raman, Nikhil Gupta, and Yogananda Jeppu. Framework for formal verification of machine learning based complex system-of-systems. *Insight*, 26(1):91–102, 2023.

[148] Carl Rasmussen and Zoubin Ghahramani. Occam's razor. *Advances in neural information processing systems*, 13, 2000.

[149] Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.

[150] LS Robson, HS Shannon, LM Goldenhar, and AR Hale. Quasi-experimental and experimental designs: more powerful evaluation designs. *Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries. Department of Health and Human Services: Cincinnati, OH*, pages 29–42, 2001.

[151] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[152] Peter H Rossi, Mark W Lipsey, and Gary T Henry. *Evaluation: A systematic approach*. Sage publications, 2018.

[153] PH Rossi and HE Freeman. *Evaluation: A systematic approach (pp. 375-415)*. Newbury Park, CA: Sage, 1989.

[154] Donald Rubin. Estimating causal effects of treatments in experimental and observational studies. *Ets research bulletin series*, 1972(2):i–31, 1972.

[155] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[156] Donald B Rubin. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

[157] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.

[158] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.

[159] Donald B Rubin. Causal inference through potential outcomes and principal stratification: application to studies with "censoring" due to death. *Statistical Science*, pages 299–309, 2006.

[160] Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.

[161] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[162] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7), 2018.

[163] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.

[164] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[165] Bertrand Russell and Alfred North Whitehead. *Principia Mathematica*. Cambridge University Press, 1910.

[166] John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society*, pages 393–415, 1958.

[167] Henry Scheffe. *The analysis of variance*. John Wiley & Sons, 1999.

[168] Daniel P Scheitrum, Colin A Carter, and Cesar Revoredo-Giha. Wti and brent futures pricing structure. *Energy Economics*, 72:462–469, 2018.

[169] Erwin Schrödinger. Die gegenwärtige situation in der quantenmechanik. *Naturwissenschaften*, 23(50):844–849, 1935.

[170] M Scriven. The pathway comparison model of evaluation, 1972.

[171] Michael Scriven. Prose and cons about goal-free evaluation. *Evaluation Comment*, 3(4).

[172] Michael Scriven. The methodology of evaluation, social science education consortium. publication 110, . 1966.

[173] Michael Scriven. Maximizing the power of causal investigations: The modus operandi method. *Evaluation studies review annual*, 1:101–118, 1976.

[174] Michael Scriven. *Evaluation thesaurus*. Sage, 1991.

[175] Michael Scriven. Evaluation as a discipline. *Studies in Educational Evaluation*, 20(1):147–166, 1994.

[176] Michael Scriven. The logic of evaluation. *Dissensus and the search for common ground*, 1:1–16, 2007.

[177] Michael Scriven. Roadblocks to recognition and revolution. *American Journal of Evaluation*, 37(1):27–44, 2016.

[178] Jenn W Sellers, Camelia M Mihaescu, Kassa Ayalew, Phillip D Kronstein, Bei Yu, Yang-Min Ning, Miguel Rodriguez, LaKisha Williams, and Ni A Khin. Descriptive analysis of good clinical practice inspection findings from us food and drug administration and european medicines agency. *Therapeutic Innovation & Regulatory Science*, 56(5):753–764, 2022.

[179] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[180] Ray J. Solomonoff. A formal theory of inductive inference. parts i and ii. *Information and Control*, 7(1–2):1–22, 224–254, 1964.

[181] SPEC. SPEC CPU Benchmark Suite. https://www.spec.org/benchmarks.html#cpu.

[182] SPEC. SPEC CPU2017, 2017. Available at https://www.spec.org/cpu2017.

[183] Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[184] Robert E Stake. The countenance of educational evaluation. *Teachers college record*, 68(7):1–15, 1967.

[185] Robert E Stake. *Evaluating the arts in education: A responsiveness approach*. Merrill Publishing Co, Columbus, Ohio, 1975.

[186] Robert E Stake. Evaluating educational programmes: The need and the response. 1976.

[187] Robert E Stake. Setting standards for educational evaluators. *Evaluation News*, 2(2):148–152, 1981.

[188] Daren S Starnes, Dan Yates, and David S Moore. *The practice of statistics*. Macmillan, 2010.

[189] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[190] James Stewart. *Single variable calculus: Concepts and contexts*. Cengage Learning, 2018.

[191] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544, 2004.

[192] Daniel L Stufflebeam. Evaluation as enlightenment for decision-making. 1968.

[193] Daniel L Stufflebeam. The relevance of the cipp evaluation model for educational accountability. 1971.

[194] Daniel L Stufflebeam, Phi Delta Kappa, and Bloomington Ind. *Educational evaluation [and] decision making*. FE Peacock Itasca, IL, 1971.

[195] Daniel L Stufflebeam and George F Madaus. The standards for evaluation of educational programs, projects, and materials: A description and summary. In *Evaluation models: Viewpoints on educational and human services evaluation*, pages 395–404. Springer, 1983.

[196] DL Stufflebeam. Meta-evaluation (occasional paper no. 3). *Kalamazoo: Western Michigan University, December*, 1974.

[197] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[198] G Kasten Tallmadge. The joint dissemination review panel ideabook. 1977.

[199] Jaime A Teixeira da Silva. Citescore: Advances, evolution, applications, and limitations. *Publishing Research Quarterly*, 36(3):459–468, 2020.

[200] Jacqueline K Telford. A brief introduction to design of experiments. *Johns Hopkins apl technical digest*, 27(3):224–232, 2007.

[201] Ambler Thompson and Barry N Taylor. Use of the international system of units (si). *NIST Special Publication, Gaithersburg*, 2008.

[202] Krishnaiyan Thulasiraman and Madisetti NS Swamy. *Graphs: theory and algorithms*. John Wiley & Sons, 2011.

[203] Bruce A Thyer. *Quasi-experimental research designs*. Oxford University Press, Oxford, UK, 2012.

[204] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: System description. In *International joint conference on automated reasoning*, pages 292–297. Springer, 2006.

[205] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer*, pages 23–65. Springer, 2007.

[206] Ralph W Tyler. *Basic principles of curriculum and instruction*. University of Chicago Pres, 1950.

[207] Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of psychology, second edition*, 2, 2012.

[208] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[209] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[210] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[211] Juan D Velásquez and Vasile Palade. A knowledge base for the maintenance of knowledge extracted from web data. *Knowledge-Based Systems*, 20(3):238–248, 2007.

[212] Chenxi Wang, Lei Wang, Wanling Gao, Yikang Yang, Yutong Zhou, and Jianfeng Zhan. Achieving consistent and comparable cpu evaluation outcomes. *arXiv preprint arXiv:2411.08494*, 2024.

[213] Chenxi Wang, Lei Wang, Wanling Gao, Yikang Yang, Yutong Zhou, and Jianfeng Zhan. Achieving consistent and comparable cpu evaluation outcomes. *Technical Report, International Open Benchmark Council*, 2024.

[214] Alfred North Whitehead and Bertrand Arthur Russell. *Principia Mathematica*. Cambridge University Press, 2 edition, 1927. Cited for formal definitions of Variable and Function.

[215] James A Whittaker. What is software testing? And why is it so hard? *IEEE software*, 17(1):70–79, 2000.

[216] Robert Wild, Markus Nötzold, Malcolm Simpson, Thuy Dung Tran, and Roland Wester. Tunnelling measured in a very slow ion–molecule reaction. *Nature*, 615(7952):425–429, 2023.

[217] Philip Green Wright. *The tariff on animal and vegetable oils*. Number 26. Macmillan, 1928.

[218] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

[219] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental test of parity conservation in beta decay. *Phys. Rev.*, 105:1413–1415, Feb 1957.

[220] Chien-Shiung Wu and Irving Shaknov. The angular correlation of scattered annihilation radiation. *Physical Review*, 77(1):136, 1950.

[221] Hugh D Young, Roger A Freedman, and Lewis A Ford. *University physics with modern physics*. 2020.

[222] Jianfeng Zhan. Five axioms of things. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, page 100184, 2024.

[223] Jianfeng Zhan. A short summary of evaluatology: The science and engineering of evaluation, 2024.

[224] Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, Hainan Ye, Shaopeng Dai, and Zhifei Zhang. Evaluatology: The science and engineering of evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4(1):100162, 2024.

In this seminal work, Professor Jianfeng Zhan systematically differentiates intelligent life from normal objects through two defining properties: interrogation and free will. He presents a novel framework for fundamental interrogations: encompassing measurement, testing, reasoning and evaluation.

Building on this foundation, Zhan provides a formal definition of cause and effect, positing that evaluation fundamentally involves uncovering an object's effects. He rigorously formalizes the evaluation problem, along with its dual challenges—design and inverse problems—de-evaluation and introduces the groundbreaking discipline of Evaluatology. This discipline establishes a universal framework, including universal concepts, axioms, fundamental issues and methodology.

Collaborating with colleagues, graduate students, and postdoctoral researchers, Zhan et al. further develop foundational evaluation methodologies, explore novel pathways toward Strong Artificial Intelligence, and demonstrate the wide-ranging applications of Evaluatology across diverse domains.

BCPress
BenchCouncil Press
9 789887 159681
ISBN 978-988-71596-8-1