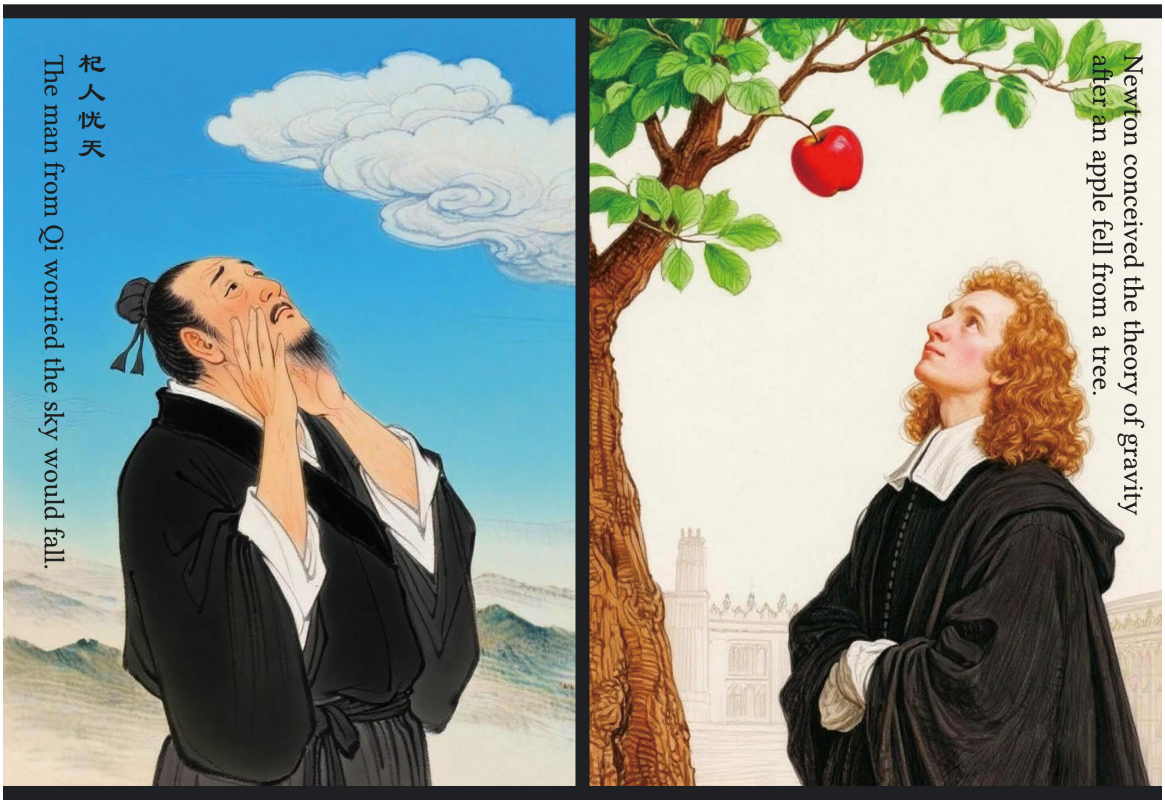
Evaluatology

The Science of Uncovering the Effects

Jianfeng Zhan, Lei Wang, Wanling Gao,
Hongxiao Li, Chenxi Wang, Fanda Fan, and Guoxin Kang



BCPress
BenchCouncil Press

Evaluatology: The Science of Uncovering the Effects

Jianfeng Zhan
Lei Wang
Wanling Gao
Hongxiao Li
Chenxi Wang
Fanda Fan
Guoxin Kang

November 23, 2025

© 2025 by Dr. Jianfeng Zhan

Published by BenchCouncil Press, Inc.

Edition 1.0, 2025

ISBN 978-988-71596-8-1

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Dedications

Dedicated to My Beloved Wife, Fei, and My Beloved Daughter, Aurora.

我曾在灿烂的星空下畅想
你
和她
会在哪里
命运
会有多少怜悯和眷顾
又会有怎样的奇遇
多少温柔的眼神
汹涌澎拜的执着
和坚毅的勇气
守护
你
我

Excerpted from my poem “纸飞机”
Dr. Jianfeng Zhan

Dedicated to My Beloved Father Zhuru Zhan and My Beloved Mother Jiuhong Chen

父亲
关于河流
关于自由
你也是热爱的
你在远方注视我四十多年
当自由变得珍贵时
我就会想起你
想起你遭受过的苦难
想起河流
想起川流不息的河流
奔腾不息
想起欢乐的鱼儿
一生的追逐
除了自由
空空如也

Excerpted from my poem "詹竹如池塘"
Dr. Jianfeng Zhan

The Starry Sky

如果
没有亘古不灭的永恒
在星空里闪耀的
是跳动的火焰
还是
扑向火焰
稍瞬即逝的星尘

Excerpted from my poem “星空”
Dr. Jianfeng Zhan

Contents

List of Figures

List of Tables

Preface: How I gave birth to Evaluatology

By Dr. Jianfeng Zhan

This morning, I woke up from a dream where I spent almost two hours writing the preface for this book. I feel very regretful. If I'd woken up earlier, I could have finished this challenging but enjoyable task.

In 2009, I was given the task of writing a technical report on information technology infrastructures for emerging computing, like Internet Services, Cloud Computing, and Big Data. At the time, my boss was Prof. Ninghui Sun, who introduced me to Prof. Kai Li, a well-known professor from Princeton University.

In the 1980s, Prof. Li graduated from the Institute of Computing Technology (ICT) at the Chinese Academy of Sciences, where I was an Assistant Professor since 2002, an Associate Professor since 2004, and later promoted to a Full Professor in 2012.

Kai impressed me in three key ways. First, when discussing innovations, he always starts by calculating the cost using current best practices—I found this truly amazing. In China, many scientists often see cost calculation as boring or even pointless, so his approach stands out.

Second, Kai has created several highly influential benchmark works. One is PARSEC, a well-known CPU benchmark. Another is ImageNet, which he co-developed with Professor Feifei Li. The AI community widely credits ImageNet as one of the key drivers behind the AI boom.

Lastly, Kai is incredibly successful in business. His startup, DataDomain, was acquired by EMC for nearly one billion US dollars, which speaks volumes about his achievements.

I admire Kai's influence. Looking back, I believe my conversation with Kai was the starting point for both Evaluatology and this book. I define Evaluatology as "the science of uncovering the effects of everything." In this book, I use the same methodology to trace the people and events that influenced both the book's creation and me while I was writing it.

During this journey, two things stand out—one was a stroke of luck, and the other was a bit unfortunate.

The first thing I want to share is BigDataBench, our first influential benchmark for Big Data. I worked on it in 2013 with Mr. Lei Wang, who was my Ph.D student at the

time. Surprisingly, we finished this work in just two weeks.

Eventually, our article was accepted by HPCA 2014. I remember clicking the submission button to upload our paper to the HPCA conference while sitting in the flight cabin. My hands were shaking. Five minutes later, my flight took off from San Francisco to Beijing.

We were very lucky—our paper received a high score, and the industry chair of HPCA was incredibly supportive and encouraging. Dr. Zhen Jia, my previous Ph.D. student, later told me that when a famous professor asked Prof. Kai Li to recommend a big data benchmark, Kai recommended our BigDataBench.

At the same time, I submitted another OS-related article to top conferences like ASPLOS, SOSP, and OSDI. But this time, we spent six years without the article being accepted. I still remember one year at ASPLOS, where one reviewer even gave us a "strong accept," which is quite rare in computer systems conferences. In the end, I decided not to publish this article.

The second thing I want to share is AIBench, another benchmark project focused on AI. I worked on it with Miss Wanling Gao, who was also my Ph.D student at the time. She's a very smart and hardworking person.

Wanling submitted the AIBench paper to the HPCA 2020 conference. She hadn't slept for almost three days before she finally clicked the submission button. I felt confident about the paper's chances, and it turned out we received very high scores: four accepts and one weak reject.

However, I noticed something unusual. The reviewer who gave a weak reject mentioned that MLPerf was already enough, and there was no need for AIBench. MLPerf is a collaborative project involving major U.S. companies and universities, making it a competitor to our work.

I decided to write an email to the industry chair, thanking him for the effort and reminding them that there's a competition between AIBench and MLPerf. I also emphasized that having two independent benchmarks would benefit the entire community.

The chair never replied to my email, and I couldn't help but feel that something was happening behind the scenes.

It's no surprise that our article was rejected for a non-technical reason—someone was clearly manipulating the process. I felt furious and complained to the conference organizers and high-level committees. But in the end, the decision still favored the industry chair of HPCA 2020. The chair sent me an email apologizing for making me feel unfairly treated.

I think Dr. Wanling Gao felt very depressed after this. She took several years to recover from the experience. Meanwhile, the MLPerf article got accepted about half a year later. Our article received a high score at Micro 2020 but was still rejected. Later, at PACT 2021, our article was initially deemed rejected. I wrote an email to the chair, a French scientist, and he felt our treatment was unfair. He asked another reviewer to recheck the article, and it was finally accepted.

By that time, MLPerf had become the "superstar" in the field, which made the whole situation even more frustrating.

This is a case study that showcases the toughness of our science and technical society and our guys. When you compare the outcomes (or "fates") of two articles and two young scientists under nearly identical visible conditions, it's like a quasi—experiment, which is one topic of this book.

Many unknown or known but hidden factors dominate the outcomes, which is what we have to face. I'm no longer angry about it. Instead, I've learned to focus on more meaningful work to overcome these struggles. One such work is Evaluatology, a fascinating and thought-provoking field I'm now pursuing.

In 2021, I realized that benchmarks, while widely used across many fields, actually have no rigorous methodology. This is true even for famous benchmarks like ImageNet. I learned this from Prof. Kai Li's open lecture. He mentioned that when he and Dr. Feifei Li (who was a young assistant professor at the time) applied for funding, some reviewers even laughed at their idea. At that time, Kai was already a very senior and well-known professor at Princeton. Because of this, I wrote an article in my launched journal TBench to call for establishing the benchmark sciences and engineering.

But what exactly is a benchmark? My former Ph.D. student, Tang Fei, now working in a famous Chinese company, once told me that when someone asked about his research field, he felt embarrassed to say he was working on benchmarks. It seemed to him like a low-status and uninteresting area of research.

In 2022, I finally understood the link between benchmarks and evaluation. I asked myself a critical question: Why are rigorous methods like Randomized Control Trials (RCT) used to evaluate drugs, while in computer science, people still rely on empirical methods like SPEC CPU?

Everyone loves to report a CPU performance number using SPEC CPU. But my Ph.D. student, Chenxi Wang, my colleagues, Dr. Lei Wang, and Dr. Wanling Gao, proved convincingly that for the same CPU, performance numbers can vary by tens or even hundreds—showing how unreliable this method can be.

I joked: Well, reporting a CPU number won't kill anyone, unlike reporting a drug's performance. But this shouldn't be the case! We're part of a science and engineering community.

In many other areas, like university rankings, the situation is even worse. A University ranking made many young boys and girls, and even their parents, feel unhappy or even depressed. How ironic is that!

In 2023, I came up with the term Evaluatology and wrote an article titled Evaluatology: The Science and Engineering of Evaluation. In the first draft, I used the term Evaluationology, but my Ph.D student consulted a native English speaker, who suggested the shorter and more natural-sounding name Evaluatology. I happily adopted the idea.

I sent the article to many scholars, and Prof. David Lilja responded with a warm and encouraging message. He said the article was very interesting and especially appreciated my evaluation axioms. He also suggested I explore Design of Experiments (DoE) in more depth. I'm truly grateful for his insightful feedback.

I also received positive feedback from several professors, including Prof. Weiping Li

from the Civil Aviation Flight University of China, Prof. Aoying Zhou, Prof. Weining Qian, and Prof. Wei Wang from East China Normal University. Their encouragement meant a lot to me.

In 2024, we organized a workshop on Evaluatology in Guangzhou, where we discussed the science and engineering of evaluation with experts and researchers. It was a great opportunity to share ideas and learn from others.

From 2024 to 2025, I dedicated almost two years to writing this book without taking a single day off. Initially, I planned to work on it alone, but I soon realized how challenging it would be to handle everything by myself. I also understood that leading my colleagues and students to work together would make the process much valuable.

That's why I invited two of my colleagues, Dr. Lei Wang and Dr. Wanling, along with one Ph.D. student, Hongxiao Li, to join me. However, after a month, we fell behind schedule, so I asked Mr. Chenxi Wang and Dr. Fanda Fan (my postdoc) to join the team as well.

Just last month, my postdoc, Dr. Guoxin Kang, developed a strong interest in Evaluatology-based AI and put in a lot of effort. I think it's only fair to invite her to join us, too!

During the two-year process of writing this book, four events stand out as worth mentioning.

One person I know well attended my public presentation about Evaluatology and got inspired by it. He quickly wrote an article and published it in a famous magazine in a short time window. Some ideas in his work clearly came from mine, which was published one month earlier.

He felt embarrassed and sent me an email to explain two things: First, he had mentioned my work without naming me in another article. Second, he wanted to co-author a new article with me. I turned down his offer, but I didn't complain to the committees handling this issue. I didn't want to hurt his career.

Throughout my career in science and technology, I've encountered many disappointing situations. For example, I once wrote a technical report, and someone asked me if I had already published it. If I hadn't, he planned to write a book based on my report by himself alone.

I didn't like this behavior at all. That's why I added several footnotes in this book, clearly showing how my ideas were inspired by others' work. I wanted to make it clear that I respect everyone's contributions.

Second, many people don't take evaluation seriously—they think it's a "soft" field. Personally, I don't agree with this view at all. I believe Evaluatology is just as hard as design, and it might even help us create a new AI paradigm.

During a group meeting, Dr. Chunjie Luo made a very convincing point: evaluation and design are actually two sides of the same coin, the so-called dual problem. His presentation was so persuasive and compelling that it really made me think.

Third, during this process, I crossed paths with Mr. Hedong Yan. He had initially planned to join my research group as a Ph.D. student, but we didn't work well together. After a year, he left. I don't plan to share the details of our differences, but I do want to

thank him for one thing—he provided many valuable references for Part IV, even though he didn't contribute directly to the work.

Fourth, one day in 2025, I suddenly thought: Evaluatology could be defined as the science of uncovering the effects of things. I got this idea, inspired by Dr. Judea Pearl and Dana Mackenzie, after reading their book: The Book of Why.

Last year, a well-known professor joked with me, "Why haven't you been fired by ICT, Chinese Academy of Sciences?" I could mention his name, but I won't—out of respect for him.

He explained the reason. Many scientists are busy applying for funding and gaining official recognition, like distinguished young scientists. It seems that I feel no interest in such things. His point is that I should be fired. That is one of the reasons that I am very grateful for the support from ICT.

When I moved from the Advanced Computer Systems Research Center to the Distributed Systems Research Center, both as director, ICT granted me one million yuan in research funding as an unsolicited gift—I never even had to apply. For this kind support, I am profoundly grateful to Prof. Xilin Chen, Director of ICT, and Prof. Ninghui Sun, Academician of the Chinese Academy of Engineering.

I'd like to end by sincerely thanking my family. Looking at it through the lens of Evaluatology, my wife has clearly been the person who has influenced me most over the past two decades.

We first met in October 2001 while hiking on Vigilance Mountain. A year later, we got married without the usual wedding celebrations. We embarked on a honeymoon journey to the enchanting and mystical Jiuzhai Valley, a place renowned for its breathtaking beauty. After that, we enjoy a simple life happily.

My daughter is my beloved treasure. I miss you very much. May you find joy in your life in Boston. Every life is unique, each carrying its own inherent dignity. This dignity is not defined by appearance but resides in the mind.

I also want to thank the many small animals, trees, and flowers I've cared for this past year. Whenever I felt completely drained, spending time with you—just watching and tending to your growth—brought me deep peace and renewal.

Part I

Contributions, and Related Work

Chapter 1

What are the Contributions of This Book

In this chapter, I will directly address the core question: what are the key contributions of this book? Together with my colleagues—Dr. Lei Wang, Dr. Wanling Gao, Ph.D. students, Mr. Hongxiao Li, and Mr. Chenxi Wang, as well as Postdocs Dr. Fanda Fan and Dr. Guoxin Kang—we have made several novel contributions that represent significant advancements beyond prior work. Mr. Qian He made contributions to the presentation of several figures.

Throughout the remainder of this book, "I" refers to Dr. Jianfeng Zhan. When referring to my colleague—whether a Ph.D. student or a PostDoc—I will use their name. The term "we" denotes multiple contributors, with details provided in the context.

1.1 Uncovering the Essence of Evaluation

I reveal that the essence of evaluation is to infer the effect of an (evaluated) object.

Gravitation between objects manifests as a universal effect; combustion embodies oxygen's effect; heredity arises from the effect of genetic and other material factors; a criminal act invariably yields a detrimental effect; law implementation generates societal effects; while CPUs and Large Language Models in computing produce objectively quantifiable effects.

To systematically reveal these diverse effects, a universal conceptual framework, model, and methodology can be established. This constitutes the primary motivation for my proposal of Evaluatology as a new discipline.

1.2 Formalization of Evaluation and its Dual and Inverse Problems

Evaluation and design are dual problems. An object and external essential objects induce overall effects on an affected object. The evaluation is to uncover the true effect of a

specific object (a cause) from the overall effect, while the design of an object aims to search for a specific object configuration to achieve the optimal overall effect.

We observe a phenomenon that consists of many objects, showing different quantities. The inverse problem of the evaluation, which I call de-evaluation, is to trace back the objects and their observed quantities to the causes, external essential objects, and their effects on the affected objects.

I will formally define those concepts.

1.3 Formal Introduction of the Discipline of Evaluatology

I coined the term Evaluatology [1] to cover this exciting science. I formally define Evaluatology as the science of uncovering the effects.

I propose the fundamental components of Evaluatology.

1. universal evaluation concepts.
2. five axioms of evaluation.
3. categories of evaluation problems.
4. fundamental issues of Evaluatology.
5. fundamental Evaluatology methodology [2].

1.4 Why the Past Efforts Failed to Establish the Discipline of Evaluation?

I systematically explain why the past efforts failed to establish the discipline of evaluation.

First, the consensus of the evaluation community referred to evaluation as "the process of determining the merit, worth, or value of things" [172, 174, 175, 176, 55, 68, 85, 196, 195]. However, this definition fails to capture the essence of evaluation, as the terms "merit," "worth," and "value" are inherently subjective—their interpretation varies significantly across individuals, which constitutes a fundamental flaw in the definition.

Secondly, past works failed to propose universal evaluation concepts, problem statements, axioms, mathematical formulations, and methodologies. Instead, the community or leading scholars rely on the encyclopedic approach to define several hundred concepts or terms as shown in [174, 118].

[1] I originally coined the term "Evaluationology". Mr Hongxiao Li suggested I change it to Evaluatology after consulting with his English teacher. I adopted his suggestion.

[2] Dr. Wanling Gao, Mr. Chenxi Wang, and Dr. Lei Wang also contributed to this methodology.

1.5 New World Model: What Distinguishes Intelligent Lives from Normal Objects?

I propose foundational assumptions and models of the three worlds, upon which a distinct pathway to achieving strong AI becomes feasible.

An object is a class of entities owning a set of properties. The world consists of objects. I propose that intelligent Lives have two unique properties that distinguish them from other normal objects: *free will* and *interrogation*.

Free will is the capacity and capability to make free and intentional choices. Free will has different degrees. *Interrogation* is the capacity and capability to understand objects and their mutual effects. Interrogation has different complexities.

I formally define the cause and effect. For objects A and B, when measurable or testable differences occur in B depending on the presence or absence of A, A is the *cause*, B is the affected object, and the measurable or testable difference in B is the *effect* on B induced by A. The *effect mechanism* is the way the cause induces the effect on the affected object.

A microscopic object world consists of microscopic objects at the scale of atoms and subatomic particles, which is governed by the principles of quantum mechanics, dominated by probability.

A normal object world consists of normal objects, determined and governed by the principle of cause and effect.

A free will world consists of the normal objects and the intelligent lives, and a free will world is governed not only by the principle of cause and effect but also by free will. That is to say, an intelligent life not only receives the effects of causes, but also has the capacity and capability to make free and intentional choices

1.6 New Interpretation of Fundamental Interrogations: Measurement, Testing, and Reasoning

I posit that comparison constitutes a primitive form of interrogation, upon which intelligent life has built three fundamental interrogations: measurement, testing, and reasoning. I offer a novel interpretation of them with Dr. Lei Wang and Dr. Fanda Fan.

In Metrology, a unit of measurement like the meter was initially defined by referencing a specific Earth-based length, later refined using light speed as the reference for enhanced precision.

Similarly, in testing, the ground truth of objects constitutes the test oracle. For instance, when an image objectively depicts a cat, algorithms are tested on diverse feline images by comparing outputs against the test oracle to determine their validity. Modern artificial intelligence has evolved precisely through this paradigm.

Reasoning, as an intellectual activity utilizing inference rules that perfectly pass such tests, can supplant physical-world operations—all underpinned by comparison.

I also propose the term Testology to cover the science of testing and its application. The motivation is to present universal testing principles and methodologies across different areas, like verifying a Physics theory, hypothesis testing in statistics, and artifact testing, such as software and hardware testing.

1.7 Revealing the Essence of Value

I interpret the value of an object as its derived effect on the stakeholder. A stakeholder is an intelligent life or an organization that consists of intelligent lives. As a stakeholder can make free and intentional choices on different objects, a value function could be formulated on the basis of different quantities of their derived effects.

1.8 Fundamental Roles and Interrelationship of Four Interrogations

I clarify the unique fundamental roles and interrelationships of four interrogations: measurement, testing, reasoning, and evaluation. There are other interrogations, which I will explain in another book.

In the epistemic hierarchy of Evaluatology, measurement, testing, reasoning, and evaluation represent four fundamental interrogations through which intelligent lives explore the unexplored world and their unknown lives, and build massive knowledge systems.

Measurement answers "how much", attributing values to countable quantities of objects; testing answers "whether", determining conformity to the test oracle through verification and falsification; evaluation answers "why" in terms of how an object influences another one; reasoning answers "why" in terms of the underlying logical mechanisms that connect causes to their effects. Together, these four interrogations form a complete cognitive cycle—from observation, to validation, to explanation [3].

Mr. Chenxi Wang, Dr. Lei Wang, Dr. Wanling Gao, Dr. Fanda Fan, and I provide many examples to support our propositions or models.

1.9 The Summary of Other Fundamental Evaluation Methodology

My collaborators systematically summarize the other fundamental evaluation methodologies. Mr. Chenxi Wang contributed to the design of experiments, randomized control trials. Mr. Hongxiao Li contributed to quasi-experiments, structural causal models, Dr. Lei Wang contributed to structural equation models and instrumental variables, and Dr. Wanling Gao contributed to the potential outcome theory.

[3] Dr. Fanda Fan's work provided a basis for this passage.

I joined with all collaborators to discuss what is the basic concepts, problem statements, assumptions, and principles of each methodology. I contributed to the part of the content.

1.10 Formal Definitions of Benchmarks and Testbed

Benchmarks and testbeds are widely used in engineering without formal definitions and rigorous methodologies. I formally established the operational definitions of benchmarks and testbeds. Dr. Wangling Gao joined me to propose the principles and methodology for the testbed.

1.11 New Possible Paths to Strong Artificial Intelligence

I conceived the core idea of the paths to strong artificial intelligence based on Evaluatology. Dr. Guoxin Kang and Dr. Wanling Gao joined me to elaborate on those ideas.

1.12 Applications of Evaluatology in Different Areas.

Many collaborators actively utilize Evaluatology in different areas, demonstrating the power of Evaluatology. In this book, Dr. Fanda Fan and I showcase how to utilize Evaluatology to evaluate science and technology research institutes.

Chapter 2

Evaluation: Ancient Practice, Undeveloped Discipline

In this chapter, I will present the state-of-the-art and state-of-the-practice of evaluation. Section 2.1 will explain why evaluation is an ancient practice, incorporating the remarks of Dr. Michael Scriven. Section 2.2 presents different evaluation concepts and theories, and ad hoc practices in different domains. Section 2.3 provides an answer to why the past efforts failed to establish the discipline of evaluation. Section 2.4 presents the summary of this chapter.

2.1 Evaluation Is an Ancient Practice

Scriven [174] defined the evaluation as "the process of determining the merit, worth, or value of things, or the result of that process" [172, 174, 175, 176]. Based on this definition, he argued that evaluation is an ancient practice older than other human practices. I incorporate his remarks in this section.

Scriven [174] posited that Logic, closely analogous to reasoning elucidated in Chapter 6, and evaluation constitute "two foundational tool disciplines with pervasive applications across diverse academic domains." He concluded that the practical utilization of informal logic and evaluation predates the establishment of any formal academic disciplines or their early forms.

According to Scriven [174], informal logic and grammar co-evolved with any language that existed before formal academic disciplines, as both disciplines fundamentally rely on linguistic structures as their bedrock for further advancement. Nevertheless, evaluation practices endure even before the emergence of linguistic capabilities, owing to their indispensable role in the creation of early artifacts. For instance, the earliest recorded craftsmen—the stoneworkers—exhibit a consistent trajectory in material quality and design sophistication, a phenomenon discernible not only at individual archaeological sites but also across millennia of human history.

In Parts II and III, I define interrogation as the capacity, capability, and process to understand objects and their mutual effects. I systematically analyze four fundamental interrogations: measurement, testing, reasoning, and evaluation.

I defined the evaluation as uncovering the effects and derived effects of an (evaluated) object. If an object has a stakeholder, I interpreted determining the merit, worth, or value of an object [172, 174, 175, 176] as uncovering the derived effect of an evaluated object on the stakeholder. For any intelligent life, not just humans but also animals, evaluation is one of the fundamental interrogation abilities, as most of them can know and predict the effects of some causes. For example, a deer knows the emergence of a lion will endanger its life. For this reason, I concluded that evaluation practices endure much earlier than the period mentioned by Dr. Michael Scriven, even in the period when human beings did not exist.

2.2 Evaluation Concepts, Theories, and Ad Hoc Practices

In this section, I will summarize different evaluation concepts, theories, and ad hoc practices in different domains.

Stufflebeam [196] suggested eight questions to be addressed in any attempt to conceptualize evaluation, based on which Nevo [128, 129] extended to ten major dimensions in a conceptualization of evaluation. I propose nine dimensions, some of which overlap with those of Nevo [128, 129], incorporating the discussions by Scriven [175].

2.2.1 Definitions

Educational evaluation pioneer Ralph Tyler [206] perceives evaluation as "the process of determining to what extent the educational objectives are actually being realized." According to my definition, this definition could be reformulated as *the process of uncovering the effect of any object in the education process to determine whether it meets the intended effect.*

Another widely accepted definition of the evaluation has been that of providing information for decision-making, suggested by various leading evaluators such as Cronbach [44], Stufflebeam [194], and Alkin [2]. According to my definition, this definition could be reformulated as *uncovering and reporting the effects of an object for decision making.* Scriven referred to evaluation as "the process of determining the merit, worth, or value of things, or to the result of that process" [172, 174, 175, 176]. Evaluators and researchers in social sciences reached a considerable consensus regarding the definition of evaluation as the assessment of merit or worth [55, 68, 85, 196], or as an activity comprised of both description and judgment [74, 184]. A joint committee on standards for evaluation, comprised of 17 members representing 12 organizations associated with educational evaluation, recently published their definition of evaluation as "the systematic investigation of the worth or merit of some objects" [195].

According to my definition, anything or the result of any evaluation process could be an evaluated object. The essence of determining the merit, worth, or value of a thing is to uncover its derived effect on the stakeholder. Alternatively, it is feasible to compare the effects or derived effects of different evaluated objects of the same class. As analyzed in Chapter 10, I consider comparison as the most primitive interrogation.

Some groups or scholars rejected the judgmental definition of evaluation. For example, the Stanford Evaluation Consortium group defined evaluation as "systematic examination of events occurring in and consequent of a contemporary program -an examination conducted to assist in improving this program and other programs having the same general purpose" [45]. According to my definition, the events occurring in and consequent to a contemporary program could be an evaluated object. Uncovering its effect naturally helps improve the program.

Cronbach and his associates [45] perceived the evaluation as "an educator [whose] success is to be judged by what others learn", rather than a "referee [for] a basketball game", who is hired to decide who is "right or wrong." According to my definition, what other learn is one of the effects of an educator; the effects of different educators could be compared.

Rossi et al. present the concept framework in their famous book [152]. Throughout the book, the terms "evaluation", "program evaluation", and "evaluation research" are used interchangeably. Although they focus on the evaluation of the social program, they claim that the evaluation research is not limited to that arena [152].

Rossi et al. [152] defined program evaluation as "the application of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action to improve social conditions." In this definition, the social programs, also referred to as social interventions, cover human services programs in the domain of "health, education, employment, housing, community development, poverty, criminal justice, and international development."

According to my definition, the social intervention programs are the evaluated object. After uncovering the effects of different programs, we can investigate their effectiveness.

2.2.2 Essences or Views

In [175], Scriven summarized several different thoughts on the essences or views of evaluations, including A: "strong decision support," B: "weak decision support," C: "relativistic," D: "rich description," E: "social process," F: "constructivist" or "fourth generation." Actually, Scriven used the term "models of evaluation." I would rather use essence or view than a model for two reasons: the essence or view is more accurate; the evaluation model has other uses in Evaluatology.

A: "strong decision support" view

View A was exemplified in, but not made explicit by, the work of Ralph Tyler, and extensively elaborated in the CIPP (Context, Input, Process, and Product) model of evaluation [194]. Unfortunately, this view does not reveal the essence of evaluation. Instead, it focuses on the purpose of the evaluation: "explication of the use of program evaluation as part of the process of rational program management, conceiving of evaluators as doing investigations aimed at arriving at evaluative conclusions designed to assist the decision-maker."

B. The "weak decision support" view

This point of view is represented by evaluation theorists such as Marv Alkin [2], who define evaluation as "factual data gathering in the service of a decision-maker who is to draw all evaluative conclusion." Similarly, this view does not reveal the essence of evaluation. Instead, it focuses on a sub-process of the evaluation: it provides decision-relevant data, and even stops short of drawing evaluative conclusions.

C. The "relativistic" view

This view is from two social scientists and essentially represents this approach [153]. It holds that evaluation should be done by using the client's values as a framework, without any judgment by the evaluator about those values or any reference to other values. Unfortunately, it does not explain what value is. In Evaluatology, if an evaluated object has a stakeholder, the value is interpreted as the derived effect of the evaluated object on the stakeholder. Please note that we have a formal definition of a stakeholder. Additionally, the client is only one stakeholder, which can not justify excluding the other clients.

D. The "rich description" approach

This view has been very widely supported–by Bob Stake [186], the North Dakota School, many of the UK theorists, and others. It claims that evaluation can be done as "a kind of ethnographic or journalistic enterprise, in which the evaluators report what they see without trying to make evaluative statements or infer to evaluative conclusions– not even in terms of the client's values (as the relativist can)."

I would like to interpret "done as a kind of ethnographic or journalistic enterprise" as a kind of interrogation. However, this view is very vague without explaining what the valid interrogation methodologies are, and how measurements or testing are performed under different interrogation conditions.

E. The "social process" school

By a group of Stanford academics led by Lee Cronbach, referred to here as C&C (for Cronbach and Colleagues [45], this view denied the importance of functions of evalua-

tion, "(i) as providing support for external decisions about programs, or (ii) to ensure accountability."

This view emphasizes denying the importance of the functions of evaluation; however, it fails to reveal the essence of the evaluation.

F. The "constructivist" or "fourth generation" approach

By Egon Guba and Yvonna Lincoln [75], it rejects evaluation as a search for quality, merit, worth, etc. Instead, they claim the evaluation outcome is "the result of construction by individuals and negotiation by groups."

This view ignores that the effect of any object is objective. It seems that they try to explain how a value (quality, merit, worth, etc) function is assigned to an object. In Evaluatology, the effect and the derived effect are both objective. The value is interpreted as the derived effect on the stakeholder.

G: A "transdisciplinary" view

Scriven held a transdisciplinary view [174] to treat evaluation as a tool discipline.

Scriven claimed that this view has four characteristics that distinguish it from the previous works [175].

First, it is "an objectivist view of evaluation, like A, holding that the evaluation is to determine the merit or worth of, for example, programs, personnel, or products." Unfortunately, when it comes to worth, merit, or value without an objective interpretation, it's easy to fall into the quagmire of subjectivity.

Second, the approach here is "a consumer-oriented view rather than a management-oriented (or mediator-oriented, or therapist-oriented) approach to program evaluation–and correspondingly to personnel and product evaluation, etc." From the perspectives of Evaluatology, the consumer, management, meditator, or therapist are different stakeholders; different views do not contradict. Instead, the effects or derived effects on different stakeholders could be inferred using the same methodologies.

Third, the approach here is "a generalized view." It is "not just a general view; it involves generalizing the concepts of evaluation across the whole range of human knowledge and practice." Unfortunately, Scriven did not propose general concepts, terminology, problem statements, axioms, mathematical notation, formulations, and methodologies. Instead, he still relies on the encyclopedic approach to define several hundred concepts or terms in [174], just as other evaluation communities did in [118].

(IV) The transdisciplinary view is "a technical one." Unfortunately, Scriven fails to find a suitable technique language, like mathematics, to define the different categories of evaluation problems, and propose universal methodologies to address different evaluation problems.

2.2.3 Function

Scriven [172] was the first to suggest the distinction between "formative evaluation" and "summative evaluation," referring to two major roles or functions of evaluation, although he was not the first to realize the importance of such a distinction.

Referring to the same two functions, Stufflebeam [193] suggested "the distinction between proactive evaluation intended to serve decision-making, and a retroactive evaluation to serve accountability." Thus, evaluation can serve two functions, the "formative" and the "summative."

Robert E. Stake [184] distinguished the distinctions between formative and summative evaluation in an analogical manner from perspectives of different stakeholders: "When the cook tastes the soup, that is formative; when the guests taste the soup, that is summative." In its formative function, evaluation is used for the improvement and development of an ongoing activity (or program, person, product, etc.). In its summative function, evaluation is used for accountability, certification, or selection.

There are other discussions on the functions of evaluation from different perspectives.

Process evaluation focuses on "the activities and events during a program or intervention, investigating why and how a program or intervention achieves its results through documenting and collecting data [107, 118]."

Impact evaluation or assessment focuses on "the outcomes or impacts of an evaluand, e.g., a program, intervention, policy, organization, or technology, aiming to make a causal inference that connects the evaluand (the evaluated object) with the outcomes [107, 118]."

From the perspective of Evaluatology, there is only one unique function of evaluation: revealing the effect or derived effect of an (evaluated) object. From this angle, for formative evaluation or process evaluation, the evaluated object is the intermediate one in the different phases of creating an artifact, or the whole process of creating an artifact. While for summative evaluation or impact evaluation, the evaluated object is the object delivered. Though the evaluated objects differ, the methodology remains the same.

2.2.4 Role

In [174], Scriven considered evaluation as one of the most powerful and versatile of the "transdisciplines"—tool disciplines such as logic, design, and statistics. He claimed "Science itself is only distinguishable from pseudoscience by means of evaluation, by evaluation of the quality of evidence, research designs, instruments, interpretations, and so on [174]."

I agree with Scriven on the fundamental role of evaluation. I thought evaluation is one of the fundamental interrogation methodologies, just like measurement, testing, and reasoning. However, Scriven falsely classifies testing into the evaluation. Distinguishing science from pseudoscience is the fundamental role of testing, as we discussed in Chapter 5.

2.2.5 Collected Data

Previous work has extensively discussed how to collect data without deeply thinking about the different natures of the evaluated object, and hence, their thoughts are ad

hoc.

For example, Stufflebeam's CIPP Model [193] suggested that evaluation focuses on "four variables for each evaluated object: (a) its goals, (b) its design, (c) its process of implementation, and (d) its outcomes." According to this approach, an evaluation of an educational project, for example, would be "an assessment of (a) the merit of its goals, (b) the quality of its plans, (c) the extent to which those plans are being carried out, and (d) the worth of its outcomes."

Stake [184] in his Countenance Model suggested that two sets of information be collected regarding the evaluated object: descriptive and judgmental. The descriptive set should focus on "intents and observations regarding prior conditions that may affect outcomes, transactions, process of implementation, and outcomes." The judgmental set of information comprises "standards and judgments regarding the same prior conditions that may affect outcomes, transactions, and outcomes."

Guba and Lincoln [74], expanding Stake's Responsive Education Model [185] and applying the naturalistic paradigm. Guba and Lincoln [74] suggest that the evaluator focused on "five kinds of information: (a) descriptive information regarding the evaluation object, its setting, and its surrounding conditions, (b) information responsive to concerns of relevant audiences, (c) information about relevant issues, (d) information about values, and (e) information about standards relevant to worth and merit assessments."

2.2.6 Standards and Criteria

Having to choose the standards and criteria to judge the merit and worth of an evaluated object is one of the root reasons why evaluation is considered subjective.

Some scholars went straight to ignore the judgmental nature of evaluation. Those who defined evaluation as an information collection activity to serve decision-making or other purposes [2, 44, 192] did not have to deal with the problem of choosing evaluation criteria.

Other scholars used "goal achievement" as the evaluation criterion without having justified its being an appropriate criterion [206, 145]. They ignored the issue of evaluation criteria.

Several attempts have been made in recent years to develop standards and criteria for evaluations of educational and social programs [39, 195, 194, 198, 136]. Even though some scholars [45, 187] have criticized the rationale for the whole standard-setting effort as being premature at the present state of the art in evaluation, there seems to be a great deal of agreement regarding their scope and content.

Boruch and Cordray [22] analyzed six sets of such standards. They concluded that there has been a large degree of overlap and similarity among them. The Joint Committee on Standards for Educational Evaluation [195] developed and published the most elaborate and comprehensive set. Chaired by Dr. Daniel Stufflebeam, these standards committees consist of a committee of 17 members, representing 12 professional organizations associated with educational evaluation. The proposed 30 standards were divided into four major groups: "utility standards ensure that the evaluation serves practical

information needs; feasibility standards ensure that the evaluation is realistic and prudent; propriety standards ensure that the evaluation is conducted legally and ethically; accuracy standards ensure that the evaluation reveals and conveys technically adequate information."

Most evaluation experts seem to agree that the criterion (or criteria) to be used for the assessment of a specific object must be determined within the specific context of the object and the function of its evaluation. With different levels of acceptance, the evaluation criteria suggested by the literature include: identified needs of actual and potential clients [195, 138, 172], ideals or social values [74, 85], known standards set by experts or other relevant groups [74, 184], or the quality of alternative objects [85, 172].

Rossi et al. [152] discussed the criteria or standard for program performance, which may manifest in different forms for various dimensions of program performance. The criteria or standards could be "the needs or wants of the target population, stated program goals and objectives, professional standards, customary practice, norms for other programs, legal requirements, ethical or moral values, social justice, equity, past performance, historical data, targets set by program managers, expert opinions, pre-intervention baseline levels for the target population, conditions expected in the absence of the program (counterfacutal), cost or relative cost." The effectiveness of a social program is gauged by the change it produces in outcomes that represent the intended improvements in the social conditions it addresses.

In Chapter 10, I posit that comparison is the most primitive interrogation, on which the principle and methodology of measurement and testing rely. From this perspective, comparing has nothing to do with subjectivity, as measurement and testing are considered objective. According to Evaluatology, we can create reference evaluated objects and compare the evaluated object to the reference one.

2.2.7 Process

The evaluation process is ad hoc and differs according to the different views of evaluation.

A theoretical approach perceiving evaluation as an activity intended to determine whether goals have been achieved [206] might recommend the following evaluation process: "(a) stating goals in behavioral terms, (b) developing measurement instruments, (c) collecting data, (d) interpreting findings, and (e) making recommendations."

According to Stake's Countenance Model [184], the evaluation process should include: "(a) describing a program, (b) reporting the description to relevant audiences, (c) obtaining and analyzing their judgments, and (d) reporting the analyzed judgments back to the audiences."

Later on, in his Responsive Evaluation Model Stake [185] suggested a continuing "conversation" between the evaluator and all other parties associated with the evaluand. He specified 12 steps of dynamic interaction between the evaluator and his audience in the process of conducting an evaluation.

Provus [145] proposed "a five-step evaluation process, including (a) clarification of the program design, (b) assessing the implementation of the program, (c) assessing

its in-term results, (d) assessing its long-term results, and (e) evaluating its costs and benefits."

The Phi Delta Kappa Study Committee on evaluation [194] presented a three-step evaluation process. It included "(a) delineating information requirements through interaction with the decision-making audiences,(b) obtaining the needed information through formal data collection and analysis procedures, and (c) providing the information to decision-makers in a communicable format."

Scriven [170] has suggested nine steps in his Pathway Comparison Model. Guba and Lincoln [74] suggest that a naturalistic responsive evaluation should be implemented through a process including the following four stages: "(a) initiating and organizing the evaluation, (b) identifying key issues and concerns,(c) gathering useful information, and (d) reporting results and making recommendations."

Rossi et al. [152] considered that the evaluation of a program "generally involves assessing five domains: the need for the program; its design and theory; its implementation and service delivery; its outcome and impact; and its efficiency."

The evaluation process in Evaluatology has a universal process, regardless of the evaluated objects.

2.2.8 Methodologies

In Part IV, we summarize other widely used evaluation methodologies. They include Design of Experiments (DoE), RCTs, instrumental variables, potential outcomes, quasi experiments, structural causal models, and structural equation models. Uninformatively, there lack of a systematic evaluation of those evaluation methodologies, which belong to the category of meta-evaluation.

2.2.9 Ad Hoc Practices in Different Domains

This subsection presents a concise overview of ad hoc evaluation practices in different domains, partially based on our previous work [224].

In the Field of Business

Camp [28] defines benchmarking as "the search for those best practices that will lead to the superior performance of a company." Benchmarking consists of two primary steps [28]: (1) establishes operation targets based on industry best practices; (2)"a positive, proactive, structured process leads to changing operations and eventually achieving superior performance and competitive advantage." In the study conducted by Andersen et al. [4], the essence of benchmarking is summarized as "the quest for knowledge and learning from others."

From the perspectives of Evaluatology, the process, the individual, policies, or the intermediate objects in business could be evaluated. The so-called best practices are also an evaluated object. The targets of the operations are some forms of the effects

of the evaluated objects. "A positive, proactive, structured process" is essentially the process of uncovering the effects of and performing trials on different evaluated objects, e.g., policies, to "change operations and eventually achieve superior performance and competitive advantage."

In the Field of Finance

In the fields of finance and education, indices are widely used as benchmarks to assess the overall performance of individuals or systems under study. These indices are derived by calculating the weighted average of a selected group of individuals or systems [58].

For example, stock market indices are used as benchmarks to assess the stock market's performance in the finance field. These indices are derived by calculating the weighted average of a selected group of representative stocks [58]. Some widely recognized stock market indices include the Dow Jones Industrial Average, the S&P 500, the NASDAQ Composite, and the Shanghai Stock Exchange Composite Index. Different indices employ varying calculation methods. The most common approach is the weighted average method, which determines the index value based on the weighted average of the constituent stock prices. Another method is the geometric mean method, which calculates the geometric average of the stock prices and adjusts it using a base period price. Typically, stock market indices are published at the close of each trading day. Some index providers offer real-time index data, enabling investors to stay informed about the latest market conditions.

The Brent benchmark is used to determine the price of Brent crude oil [168]. Brent crude oil is a type of light and low-sulfur crude oil produced from oil fields in the North Sea region. Due to its relatively stable supply and high quality, Brent crude oil has become a significant benchmark in the international oil market. Traders, investors, and industry participants worldwide reference the Brent benchmark to track and evaluate the price of Brent crude oil.

In finance, indexes or benchmarks are essentially the reference objects through which we compare, as we discussed in Chapter 10.

In the Field of Social Sciences:

According to Rossi et al., [152], at the earliest, Thomas Hobbes and his contemporaries tried to "use numerical measures to assess social conditions and identify the cause of mortality, morbidity, and social disorganization in the discipline of social science."

Rossi et al. [152] define program evaluation as the process of using social research methods to systematically assess programs aimed at "improving social conditions and our individual and collective well-being," to provide answers to the stakeholders. Rossi et al. [152] summarize the five domains of evaluation questions and methods that exhibit strong interplays: "(1) the need for the programs, (2) program theory and design, (3) program process, (4) program impacts, and (5) program efficiency."

From the perspective of Evaluatology, the essence of program evaluation is to uncover

the effect of the social program on the social conditions, individuals, and collective well-being.

However, from the perspectives of Evaluatology, the five domains of evaluation questions should be addressed with different interrogations, not a single evaluation. The need for the program could be interrogated by measurement. The program theory needs to be tested before implementation. The program process or its components could be evaluated to uncover their effects.

In the Field of Computer Science

Within the computer science field, there are varying viewpoints and perspectives. For example, Hennessy et al.[80] highlight the significance of benchmarks and define them as "programs specifically selected for measuring computer performance." On the other hand, John et al.[94] compile a book on performance evaluation and benchmarking without providing formal definitions for these concepts.

Kounev et al.[104] present a formal definition of benchmarks as "tools coupled with methodologies for evaluating and comparing systems or components based on specific characteristics such as performance, reliability, or security." The ACM SIGMETRICS group[26, 102] considers performance evaluation as "the generation of data that displays the frequency and execution times of computer system components, with a preceding orderly and well-defined set of analysis and definition steps."

The SPEC CPU benchmark, known as SPEC CPU [181], is widely recognized as the most renowned benchmark suite for CPU performance evaluation. Throughout its history, six versions of the SPEC CPU benchmark suite have been released, with the latest version being SPEC CPU2017 [182]. The SPEC CPU workloads cover a broad range of compute-intensive tasks.

The performance evaluation metric used in SPEC CPU is based on the execution time. The reported score of SPEC CPU represents the ratio of its execution time compared to that of a reference machine. To ensure the credibility of the results, the overall metrics are calculated as the geometric mean of each respective ratio. Each ratio is based on the median execution time from three runs or the slower of the two runs [182].

Dongarra et al. [51] proposed the LINPACK benchmark for evaluating high-performance computing (HPC) systems. The LINPACK Benchmark is designed to solve dense linear systems of equations of order n, represented by the equation $Ax = b$. It originated from the development of the LINPACK software package in the 1970s.

From the perspectives of Evaluatology, the benchmarks in the discipline of computer science are a component of simple evaluation conditions, that is, essential external objects (EXOs), in Chapter 14.3.

In the Field of Artificial Intelligence

As shown in Figure 2.1 [1], ImageNet is a significant benchmark in the field of computer vision, consisting of 14,197,122 high-resolution images manually annotated across 21,841

[1]Dr. Chunjie Luo contributed to this figure. He is one of the authors of our previous work [224].

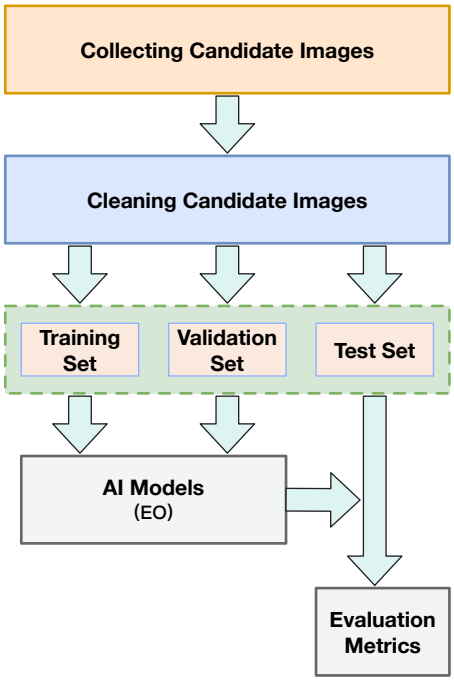


Figure 2.1: The ImageNet evaluation working process.

distinct categories, commonly known as ImageNet-21K [47]. These categories encompass a wide range of objects, animals, and scenes.

The ILSVRC (ImageNet Large Scale Visual Recognition Challenge) is an annual computer vision competition that focuses on a subset of ImageNet-21K called ImageNet-1K [164]. It aims to evaluate the performance of deep learning models in tasks such as image classification and object detection, providing specific task configurations and evaluation criteria.

ImageNet-1K is primarily used for image classification tasks and consists of 1,281,167 training images, 50,000 validation images, and 100,000 test images. The evaluation metrics commonly used in ILSVRC include Top-1 accuracy, which measures the match between the predicted category and the true category of the image, and Top-5 accuracy, which indicates if the true category of the image is among the top five predicted categories by the model.

From the perspectives of Evaluatology, the benchmarks in the discipline of artificial intelligence are a component of simple evaluation conditions, similar to those of the discipline of computer science.

In the Field of Medicine

The evaluation in the field of medicine can be traced back to the early medical eras, although there are no documented records. A rigorous modern medical evaluation methodology and system were established as early as 1938 [32]. Clinical trials, with a history

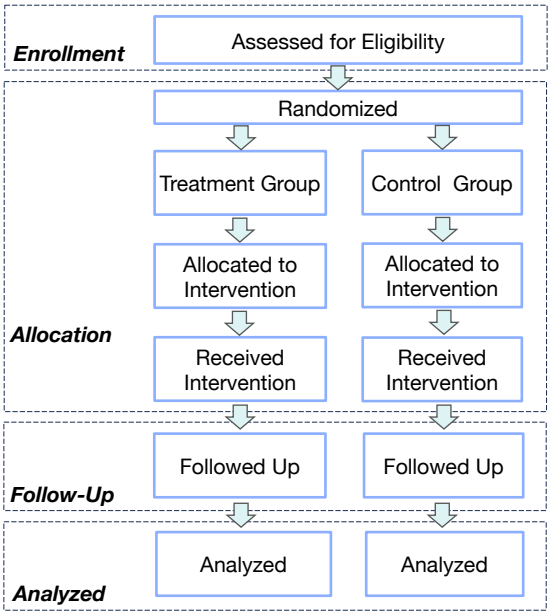


Figure 2.2: The randomized controlled trials (RCTs) evaluation process. [123]

spanning over 250 years, are the primary and widely recognized method for medical evaluation. They are defined as experimental designs to evaluate the potential impact of medical interventions on human subjects [93].

Currently, clinical trials based on experimental designs can be categorized into various types, including randomized trials, double-blind trials, prospective trials, and retrospective trials [178].

As illustrated in Figure 2.2, Randomized Controlled Trials (RCTs), considered the gold standard for medical evaluation, possess a rigorous and reliable theoretical framework [124]. However, their high time and financial costs limit their application. In addition, RCTs are difficult, if not impossible, to apply in Physics, Chemistry, or Biology to trace the causes of many effects [2].

To compensate for the shortcomings of RCTs, emerging clinical evaluation methods, such as Real-World Data assessment and digital clinical trials, have been proposed [90, 146]. These novel medical assessments are still in their early stages and have noticeable deficiencies in their theoretical foundations, such as a lack of rigor and reliability.

In the field of Psychology

In the field of psychology, social and personality psychologists often rely on scales, such as psychological inventories, tests, or questionnaires [65], to evaluate psychometric variables [65]. These variables include attitudes, traits, self-concept, self-evaluation, beliefs,

[2]This comment is inspired by Dr. Lei Wang in a talk after we had lunch together on a day in the year of 2025.

abilities, motivations, goals, social perceptions, and more [65]. Essentially, the essence of the scale is a component of simple evaluation conditions.

While these tools are commonly used, it is important to recognize that they rely on virtual assessments and self-report-style evaluations, which may introduce potential distortions.

To overcome this limitation, I suggest implementing a physical application of an EC to the evaluated object, supplemented with a variety of measurement instruments. This approach aims to provide a more objective and accurate assessment of various aspects, including attitudes, traits, self-concept, self-evaluation, beliefs, abilities, motivations, goals, and social perceptions [65], by incorporating tangible and observable data.

2.3 Why Past Efforts Failed to Establish the Discipline of Evaluation?

In [172], dated back to 1966, Scriven thought evaluation is "a logical activity which is essentially similar whether we are trying to evaluate a coffee machine or teaching machines, plan for a house or plan for a curriculum. The activity consists simply of the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings." That is the first rudimentary idea on the discipline of evaluation. Since then, Scriven has published many articles towards the goal of establishing the discipline of evaluation from the 1960s to the 2010s [172, 170, 171, 173, 174, 175, 176, 177].

In their 2010 book [56], Chinese Scholars Junping Qiu et al. discuss what scientific evaluation means in Chinese. They used a Chinese Term similar to Evaluatology. Overall, they discussed and summarized many concepts and ad-hoc methods in social sciences, education, and bibliometrics. Unfortunately, they overlooked Scriven's work and failed to notice many rigorous evaluation methodologies which we discussed in Part IV.

In a summary article, dated back to 2016, Scriven [177] claimed the discipline of evaluation is established; however, I think he overstates the situation.

In [212, 224], Wang et. al showed that the state-of-the-art and state-of-the-practice evaluation methods cannot even achieve a true evaluation outcome for a specific artifact, a computer component, like a CPU. Instead, Different areas are still using ad hoc empirical methodologies as we have discussed in Section 2.2.9.

The good news is that many stringent methodologies have been developed, like DOE, RCTs, or the potential outcome framework, which we will summarize in Part IV; however, no previous work has systematically evaluated those evaluation methodologies. In Chapter 13, we consider the meta-evaluation of different evaluation methodologies as one of the most important issues in Evaluatology.

I believe two fundamental reasons contribute to the failure of past efforts to establish the discipline of evaluation.

First, the consensus of the evaluation community referred to evaluation as "the process of determining the merit, worth, or value of things" [172, 174, 175, 176, 55, 68, 85,

196, 195]. Scriven [175] claimed that it is the value-free doctrine that prevents the development of the discipline of evaluation, that is, the science and engineering community rejected to introduce the word of value into their territory.

However, the current definition of evaluation, which is the consensus of the evaluation community, fails to uncover the essence of evaluation. The words of "merit, worth, or value" in nature are subjective—varying from different people—that is one of the root reasons that contribute to the failure.

Secondly, Scriven [174, 175, 176] envisioned that there is a core subject in the discipline of evaluation. However, the core subject of the evaluation was never articulated. The past work [174, 175, 176, 56] failed to propose universal concepts, terminology, problem statements, axioms, mathematical notation, formulations, and methodologies. For example, without those universal ones, Michael Scriven relies on the encyclopedic approach to define several hundred concepts or terms in [174], just as other evaluation communities did in [118].

2.4 Summary

This chapter overviews evaluation's current state, delving into its historical roots and theoretical foundations. It examines evaluation as an ancient practice, but an undeveloped discipline, and explores various evaluation concepts, theories, and ad hoc practices in different domains. Finally, it addresses the reasons behind past failures in establishing evaluation as a discipline.

Part II

Fundamental Interrogations

Chapter 3

Fundamental Concepts

In this chapter, I rigorously delineate six foundational concepts: object, interrogation, subject, proposition, model, and axiom system, drawing upon definitions from key references [1, 11, 190, 188, 224]. Figure 3.1 shows the relationship among the six concepts.

3.1 Object

The world consists of objects. An *object* is a class of entities owning a set of *intrinsic characteristics* that can be interrogated by those other than itself [224]. An intrinsic characteristic is often referred to as *a property*. I will formally define interrogation in Section 3.2.

A typical object could be a thing, a life, a phenomenon, an abstract concept, a process, or even a policy in the natural and social sciences, as well as engineering.

Counting is assigning a value to a property. A *quantity* is a countable property of an object, such as volume or mass. A *variable* [214] is a symbol that represents an unspecified or changing quantity.

An object has other properties that are not countable. Psychologist Stanley Smith Stevens explored this topic and classified four natures of quantities: nominal, ordinal (based on order [17]), interval, and ratio [189]. Definitely, there are other natures of quantity. For example, I consider interrogation and free will as two fundamental properties of intelligent life, which are also objects.

In the original article, Stanley Smith Stevens used the term "levels or scales of measurement." I use "nature" to avoid confusion with the specific meaning of "scales" or "levels" used in this book.

The nominal nature represents "the simplest form of measurement, where numbers serve as labels or identifiers, establishing an equality relation." The ordinal nature, in contrast, involves "ranking the items in a specific order." The interval nature exhibits an "equality of interval relation, where (1) the choice of a zero point is based on convention

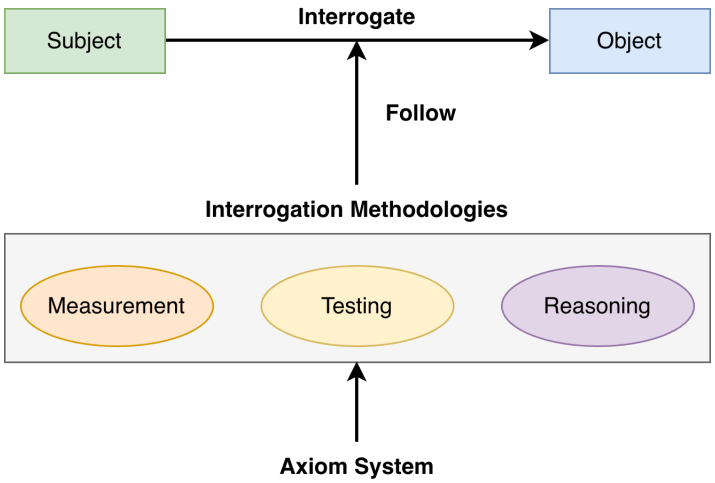


Figure 3.1: The relationships among the concepts of object, subject, interrogation, and axiom system.

or convenience, (2) there is rank ordering, and (3) the scale remains invariant when a constant is added to all values, preserving the differences between them." The ratio nature allows for "all four types of relations: equality, rank-ordering, equality of intervals, and equality of ratios."

Several Examples on Objects

A thing example

The weight of a table could be measured. A table is the object; its weight is its *quantity*.

A life example

A man could measure his height. A man is both the object and the subject; the height is his *quantity*.

A natural phenomenon example

The time of sunrise could be measured. Sunrise is the object, and the time of sunrise is its *quantity*.

An abstract concept example

A superstar can measure his or her number of social media followers. A superstar is both the object and the subject, and the number of social media followers is his or her *quantity*.

A policy example

The number of applicants applying for welfare could be measured. A welfare is the object, and the number of applicants is its *quantity*.

An *object* could be an *individual* or a *system*. An individual consists of components. A *system* is a coherent entity comprising interacting or interdependent individuals [1, 11, 224]. A system could be recursive. In the rest of this article, when referring to an object, we do not distinguish between an individual and a system unless stated explicitly.

An object has many instances. A *population* is the entire set of object instances, while a *sample* represents a smaller subset of object instances from the population [188]. A *parameter* is a number that describes some quantities of the population, while a *statistic* is a number that describes some quantities of a sample.

3.2 Interrogation

I begin by defining *interrogation* as *the capacity, capability, and process to understand objects and their mutual influences*. In Part II, I focus on the capacity, capability, and process to understand objects. The discussion of how objects influence one another will be addressed in Part III.

I define *an interrogation condition* as a setting under which objects and their mutual influences are interrogated. An interrogation condition consists of all objects that impact the interrogation outcomes. *Data* are raw interrogation outcomes or their derived ones in different interrogation conditions.

Interrogation has different levels of complexity. *Observation* is a kind of interrogation without the chance of changing the interrogation conditions. Instead, *experiment* is a kind of interrogation with the chance of changing the interrogated object, the interrogation conditions, and the interrogator.

There are three fundamental interrogations: *measurement*, *testing*, and *reasoning*. We will discuss them in Chapters 4, 5, and 6.

3.3 Subject

Free will is the capacity and capability to make free and intentional choices. Free will also has different degrees.

An automatic object is an object capable of interrogation but lacking free will, while *an intelligent life* is an object capable of interrogation with free will. Interrogation and free will are essentially two properties of an intelligent life. The *subject* is either an automatic object or an intelligent life.

A subject could be an object to be interrogated. Also, a subject could interrogate itself. An *artifact* refers to an object that demonstrates intentional conjecture, design, and fabrication by a subject.

According to the above definition, an *object*, defined in Section 3.1, could be redefined as a class of entities owning a set of *properties*, which a subject can interrogate.

Several Examples on Interrogation

A measurement example

A man could measure his height. A man is both the object and the subject; his height is the *quantity*.

A testing example

A man could test his visual acuity. A man is both the object and the subject; his visual acuity is the *properties*. The standard visual acuity chart serves as the *testing oracle*.

A reasoning example

Major Premise (Universal): All human-beings will die. *Minor Premise (Particular)*: Alice is a human being. *Conclusion*: Therefore, Alice will die.

3.4 Proposition

A *proposition* is a testable statement about an object. I will formally define what is testable in Chapter 5. A big letter A stands for a proposition.

Big letters $\mathcal{A}$ in calligraphy stand for a set that includes several (finite or infinite) numbers of propositions. $\mathcal{A}$ is usually *the power set* of A.

The power set of a given set A refers to a new set consisting of all possible subsets of A (including the empty set and A itself), denoted as $\mathcal{A}(A)$ or 2^A or $\mathcal{A}$.

A *premise* is a proposition that serves as the foundational statement from which a conclusion is logically derived. The *assumption* is a proposition accepted as truth or fact without proof, and it is a *provisional premise* adopted within a specific logical argument.

Hypothesis is a kind of proposition of an object. The difference between a hypothesis and an assumption is as follows.

Where P is a set of premises:

- *Assumptions* are elements of P: $P = \{A_1, A_2, \ldots, A_n\}$.
- *Hypotheses* are testable statements derived from P: $H \subseteq \mathcal{P}(P)$, which reads as "Let P be a set, and H is a subset of the power set of P."
- Validity holds only if: $(\bigwedge_{i=1}^{n} A_i) \to H$, which reads as "Validity holds only if the conjunction of all A_i from $i = 1$ to n implies H."

3.5 Model

A *model* is a streamlined representation of an object [221, 224]. A model can manifest as a physical, mathematical, or other construct. A valid model that passes the testing by a subject other than itself is a kind of *fact or truth*.

A mathematical model embodies a mathematical representation, frequently expressed through functions or equations, that captures the essence of an object.

Let's take the function as an example. A function [214] is a relation $f : X \to Y$ between sets X (domain) and Y (co-domain) that maps each $x \in X$ to exactly one $f(x) \in Y$.

According to [190, 224], "a function, denoted as f, is a rule that assigns a unique element, referred to as $f(x)$, from a set X to each element in a set Y." In this context, "the domain, denoted as X, refers to the set of all possible values for which the function is defined [190]." On the other hand, "the range of the function, denoted as $f(x)$, consists of all the possible values that $f(x)$ can take as x varies within the domain [190]."

The *independent variable* is represented by "a symbol that encompasses any arbitrary number within the domain of the function [190]." A *dependent variable*, represented by a symbol, "is used to denote a number within the range of the function [190]."

A random variable [103] is a measurable function $X : \Omega \to \mathbb{R}$ that extends from a probability space $(\Omega, \mathcal{F}, P)$ to $\mathbb{R}$ and assigns a number to each $\omega \in \Omega$. In this formulation, Ω is the total sample space of all possible events, $\mathcal{F}$ is the set of subsets of Ω representing events, and P stands for the probability.

For a random variable X, a distribution function [103] is a function $F_X : \mathbb{R} \to [0, 1]$ that is defined as $F_X(x) = P(X \leq x)$, stays non-decreasing, maintains right-continuity, and has $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.

3.6 Axiom System

A *fact or truth* is a proposition or model about an object that can be proven true or verified objectively by a subject other than itself. *Knowledge* contains facts or truths about objects.

The *axiom system* is *a collection of self-contained assumptions* for domain-specific knowledge or those beyond a specific domain.

For a certain proposition A, Proof(A) stands for a sequence of propositions, which starts from axioms, ends at A, and obeys the reasoning rules. The formulation $A \vdash B$ stands for the fact that there is at least a proof for B from A. If A is a false proposition, Proof(A) does not exist.

This book is built upon the axiom system. The next subsection explains what an axiom system is from different angles.

3.6.1 A Classical Perspective

From a classical perspective, axioms are self-evident assumptions.

- *Self-evident*: Axioms were seen as fundamental assumptions so obviously true that they required no proof.
- *Truths or Facts*: They were considered universal, unquestionable foundations.

A typical example of an axiom system

Euclid's 1st Axiom: "A straight line segment can be drawn between any two points."

The classical perspective, which is rooted in Euclidean geometry and rationalist philosophy (e.g., Descartes), holds that:

- *Self-evidence*: Axioms require no proof due to their intrinsic clarity. $\forall A \in \mathcal{A}_{\text{classical}}$, $\text{Proof}(A) = A$ where $\mathcal{A}_{\text{classical}}$ denotes classical axiom sets.
- *Truth or Fact*: Axioms are universally valid *a priori*:

$$\vdash A \quad \text{for all } A \in \mathcal{A}_{\text{classical}}. \tag{3.1}$$

3.6.2 Modern Perspective

Axioms are *not necessarily "self-evident" or absolute truths.* Instead, the consistency and arbitrariness of an axiom system are often used.

Consistent($\mathcal{A}$) stands for the fact that there is no proof from any propositions in $\mathcal{A}$ that leads to a contradiction. Arbitrary(A_1, A_2) stands for that, any inference taking A_1 or A_2 as an axiom is equivalent, namely $\forall B(A_1 \vdash B \iff A_2 \vdash B)$.

Axioms are defined as:

- *Foundational assumptions*: Arbitrary starting points chosen to build a logical system.
- *Defined by consistency*: Their validity depends on whether they generate non-contradictory results.
- *Relativity*: What is "self-evident" in one system may not hold in another.

A typical "self-evident" example in one system, which may not hold in another.

- *Euclid's 5th Axiom* (Parallel Postulate): "Through a point not on a line, exactly one parallel line exists."

- *Hyperbolic Geometry Axiom*: "Infinitely many parallel lines exist."

Both systems are logically consistent despite contradicting each other → Axioms are *conventions*, not universal truths.

Contemporary mathematics redefines axioms as:

1. *Formal Foundations*: $\mathcal{A}_{\text{modern}} := \{A_1, \ldots, A_n\}$ such that Consistent($\mathcal{A}_{\text{modern}}$) where consistency means no contradiction arises: $\nexists S\,(\mathcal{A}_{\text{modern}} \vdash S \wedge \neg S)$.
2. *Arbitrary Choices*: Axioms are conventions, not truths. For example:
   - Euclidean vs. non-Euclidean parallel postulates

3.6.3 Key Contrast

Classical View	*Modern View*
Axioms $\subseteq$ Truths	Axioms $\subseteq$ Assumptions
Self-evident	System-dependent

Table 3.1: Classical vs. modern view of axioms

3.6.4 Implications

Gödel's Incompleteness Theorems [62] further show:

- No consistent axiom system $\mathcal{A}$ can prove all arithmetical truths.

3.7 Interpreting Objects from a Perspective of Algebraic Structure

The algebraic system is established based on the axiomatic system. In this section, we describe objects from an algebraic perspective [1]. An algebraic structure [134] is a set A with one or more operations or mathematical properties that satisfy the specific axioms, e.g., $+ : A^2 \to A$, $\times : A^2 \to A$).

For example, a ring is denoted as $(A, +, \times)$. A ring is an algebraic structure.

[1] Mr. Hongxiao Li is the primary contributor of this section.

A category [54] is a mathematical structure consisting of a class of objects, denoted as $\mathcal{O}(\mathcal{C})$, and, for each pair of objects $A, B \in \mathcal{O}(\mathcal{C})$, a set of morphisms $\mathrm{Hom}_{\mathcal{C}}(A, B)$ from A to B.

Furthermore, a category must satisfy three axioms. First, morphisms admit composition: if $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$ and $g \in \mathrm{Hom}_{\mathcal{C}}(B, C)$, then their composition $g \circ f$ belongs to $\mathrm{Hom}_{\mathcal{C}}(A, C)$. Second, the composition of morphisms is associative: if $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$, $g \in \mathrm{Hom}_{\mathcal{C}}(B, C)$, and $h \in \mathrm{Hom}_{\mathcal{C}}(C, D)$, then the composition must satisfy the equation $h \circ (g \circ f) = (h \circ g) \circ f$. Third, each object has an identity morphism: for each $A \in \mathcal{O}(\mathcal{C})$, there exists an identity morphism $\mathrm{id}_A \in \mathrm{Hom}_{\mathcal{C}}(A, A)$ such that for all $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$ and $h \in \mathrm{Hom}_{\mathcal{C}}(C, A)$, the equations $f \circ \mathrm{id}_A = f$ and $\mathrm{id}_A \circ h = h$ hold.

A morphism [54] is a mathematical arrow (denoted as $f : A \to B$) that connects two objects A and B within a category $\mathcal{C}$. It belongs to the set $\mathrm{Hom}_{\mathcal{C}}(A, B)$ (the hom-set of the category) and can be composed with other morphisms. Specifically, if $f : A \to B$ and $g : B \to C$, their composition $g \circ f : A \to C$ satisfies the associativity condition: for any $h : C \to D$, we have $h \circ (g \circ f) = (h \circ g) \circ f$.

An object or its model can be represented as a variable, including a definite variable or a random variable, a function, a set, an algebraic structure, or other complicated structures like a category.

3.8 Summary

This chapter presents a concise system of concepts, including object, interrogation, subject, proposition, model, and axiom system.

Chapter 4

Metrology

Metrology is the science of measurement and its applications [17]. This chapter presents our interpretation of Metrology, including the basic concepts, problem statements, assumptions, methodologies, and case studies. Dr. Lei Wang contributed to Sections 4.7, 4.8.

4.1 Basic Concepts

A *reference* is a convention or standardization, or axiom, as I have discussed in subsection 3.6.2.

The *unit of measurement* is a definition of an ideal reference object with a unit quantity.

4.2 Definition of Measurement

I define *measurement* as attributing values to a quantity of an object by comparing with the unit quantity of a reference object under an interrogation condition. Measurement is a kind of experiment, as the subject could control the reference object. *Being measurable or measurability* means that an object's quantity can be compared with that of a reference object.

In metrology, measurement is often defined as "objective obtaining of one or more values attributed to a quantity [17]." I much like my definition, as it reveals the essence of measurement.

Another widespread definition of measurement in the social sciences is "the assignment of numerals to objects or events according to some rule. [189]", dating back to 1946.

4.3 Problem Statement

I state the measurement problem as follows: Given a quantity, how can we define an ideal reference object with a unit of measurement, and hierarchically realize the definition of the reference object with different accuracies and overheads?

Specifically, the input of the measurement problem is a given quantity of objects, the outputs include the definition of a unit of measurement, and hierarchical realizations of the unit of measurement. The constraints are the accuracy and overhead of the definitions and realizations.

4.4 Basic Assumptions

For measurement, there are three basic assumptions.

First, for a countable property, there is a true value, independent of any other object. The *true value of a quantity* represents an inherent property of an object that is independent of any observer. For example, it can be the radius of a specific circle or the kinetic energy of a particular particle within a given system [97, 17]. In terms of measurement, the true quantity value is a target that any measurement approaches.

Second, an ideal reference object possesses a utilizable property that can serve as the basis for defining a unit of measurement.

Third, for a given quantity of any object, the subject could realize the definition of a reference object with a unit quantity, independent of time, space, and subjects.

These assumptions have four implications as follows.

- The reference object is from the definition by the subject, which is a convention or axiom.
- A ideal reference object with a utilizable property plays an important role in defining the unit of measurement.
- The unit quantity is independent of time, space, and any subjects.
- The subject could realize a reference object with a unit quantity.

4.5 Fundamental Roles of Measurement

Measurement plays two important roles in the subjects' interrogations. First, it provides a universally available reference for the same quantity of different objects. Second, it assigns values to the same quantity of different objects, which are the basis for generating propositions or models. The latter are the essential elements in testing and reasoning.

Building the measurement system obviously relies upon reasoning, which we will discuss in Chapter 6. In metrology, four implicit assumptions, which I clarify in Section 4.4, are essentially an axiom system.

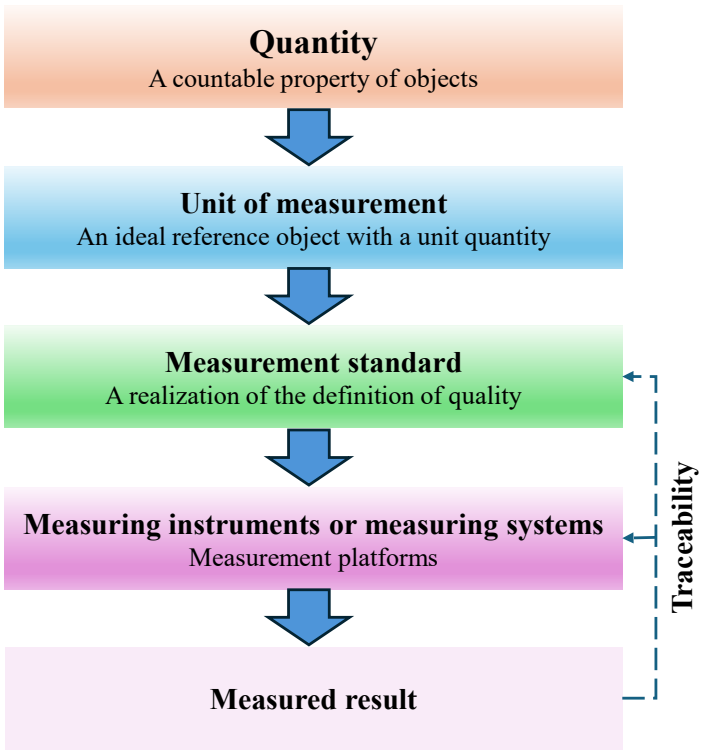


Figure 4.1: A simplified yet systematic conceptual framework for metrology [17, 97].

4.6 Methodology

The essence of metrology lies in quantities and their corresponding measurements. Figure 4.1 illustrates the systematic methodology. As shown in Figure 4.1, the quantity, the unit of measurement, the measurement standard, and measurement traceability are key components of Metrology.

The *unit of measurement* plays a fundamental role in the field of measurement [97], acting as a reference standard for comparing quantities of the same kind. It is a real scalar magnitude that is defined and adopted by convention, ensuring consistency across measurements. The unit serves to standardize measurements, allowing for their comparison and ensuring that different systems and observers can communicate precise, comparable data. Units are established through international agreements, forming the basis for uniformity and accuracy in scientific, industrial, and everyday applications.

Measurement standard [97] is a realization of the definition of quality. It is characterized by a stated metric value and an associated measurement uncertainty. To establish a measurement standard, it is important to use a *measurement methodology* that is both repeatable (performed by the same team) and reproducible (performed by different

teams). This ensures consistency and reliability in the reference for measurements. Such measurements can be conducted using *measuring instruments* or *measuring systems* [17], providing a reliable foundation for further analysis and comparison.

The hierarchy of measurement standards follows a progression from lower to upper levels, with increasing accuracy and cost. This progression starts from national measurement standards and extends to international standards. As a property of a measurement result, *measurement traceability* [17] establishes a connection between the result and a reference (measurement standards, measuring instruments, and measuring systems). This connection is established through a documented, unbroken chain of calibrations, with each calibration contributing to the measurement uncertainty. To ensure accuracy, each level of measurement standards in the hierarchy should be calibrated using a higher standard with greater precision.

4.7 Definition and Realization of the Fundamental Quantities

As mentioned earlier, a quantity is a countable property of an object that has a true value independent of any other object. The corresponding definition of a unit of measurement comes from an ideal reference object with a utilizable property. For example, length is a fundamental quantity, and its unit of measurement, the meter, is currently defined based on the speed of light in a vacuum. In this case, the light serves as the ideal reference object, and its property—the constant speed—is utilized to provide a precise and universally reproducible standard for measurement.

The international system of metrology encompasses *seven fundamental quantities*: time, length, mass, electric current, thermodynamic temperature, amount of substance, and luminous intensity [17]. These quantities form the foundation of all physical measurements, and all physical units and measurements are defined based on these fundamental quantities.

The seven fundamental quantities form the basis of the International System of Units (SI) and provide the foundation for all physical measurements. The International System of Units is defined by fixing the numerical values of seven defining physical constants as follows:

- *second* (s)
  The second, the SI unit of time, is defined by taking the fixed numerical value of the cesium frequency, $\Delta\nu_{\mathrm{Cs}}$, the unperturbed ground-state hyperfine transition frequency of the cesium-133 atom, to be $\Delta\nu_{\mathrm{Cs}} = 9\,192\,631\,770$ Hz, which defines the duration of one second [48]. In this case, the cesium-133 atom serves as the ideal reference object, and its property—the precise frequency of the hyperfine transition—is utilized to provide a standard for measuring time.

- *meter* (m)
  The meter,the SI unit of length, is defined by fixing the numerical value of the speed

of light in a vacuum, $c = 299\,792\,458$ m/s, such that one meter is the distance light travels in a vacuum during a time interval of $1/299\,792\,458$ seconds [48]. In this case, the ideal reference object is light in a vacuum, and its property of a fixed, invariant speed is used, thereby providing a precise and universally reproducible standard.

- *kilogram* (kg)
  The kilogram,the SI unit of mass, is defined by taking the fixed numerical value of the Planck constant at $h = 6.626\,070\,15 \times 10^{-34}$ kg m^2 s^{-1} [48]. In this case, the ideal reference object is a Kibble balance, a precision instrument used to measure the Planck constant. Its key property is the ability to measure the mechanical power required to balance the force of gravity on a mass using an electromagnetic force. By linking this property to the Planck constant, the Kibble balance enables the definition of the kilogram in terms of fundamental constants, ensuring a precise and reproducible standard of mass.

- *ampere* (A)
  The ampere,the SI unit of electric current, is defined by taking the fixed numerical value of the elementary charge at $e = 1.602\,176\,634 \times 10^{-19}$ C, where 1 C = 1 A s [48]. In this case, the ideal reference object is a single proton or electron. The elementary charge is the smallest indivisible unit of electric charge. The utilizable property is that the elementary charge carried by a single proton or electron is absolutely invariant and constant, which can be used to define the flow of electric charge in terms of the ampere. This precise definition of the elementary charge allows for the accurate measurement of current based on the number of charges passing a given point per second, providing a universally reproducible standard for electric current.

- *kelvin* (K)
  The kelvin,the SI unit of thermodynamic temperature, is defined by taking the fixed numerical value of the Boltzmann constant at $k = 1.380\,649 \times 10^{-23}$ kg m^{2}s^{-2}K^{-1} [48]. In this case, the ideal reference object is an ideal gas or a thermodynamic system whose particles exhibit random thermal motion. The property utilized is the average kinetic energy of the particles in the system, which is directly proportional to the thermodynamic temperature as described by the Boltzmann constant. By fixing the value of k, the kelvin is defined in terms of this fundamental physical relationship, providing a precise and reproducible standard for temperature measurement.

- *mole* (mol)
  The mole,the SI unit of amount of substance, is defined by taking the fixed numerical value of the Avogadro constant at $N_\mathrm{A} = 6.022\,140\,76 \times 10^{23}$ mol^{-1}, so that one mole contains exactly $6.022\,140\,76 \times 10^{23}$ specified elementary entities [48]. In this case, the ideal reference object is a collection of elementary entities, which can be atoms, molecules, ions, or other specified particles. The property utilized

is the exact number of these entities that corresponds to one mole. By fixing the value of the Avogadro constant, the mole is defined as the amount of substance that contains precisely this number of elementary entities, providing a universal and reproducible standard for quantifying matter.

- *candela* (cd)
  The candela, the SI unit of luminous intensity, is defined by taking the fixed numerical value of the luminous efficacy of monochromatic radiation of frequency 540×10^{12} Hz, K_{cd}. This definition implies the exact relation $K_{\mathrm{cd}} = 683\,\mathrm{cd\,sr\,kg^{-1}\,m^{-2}\,s^{3}}$ for monochromatic radiation of frequency $\nu = 540 \times 10^{12}$ Hz, where sr (steradian) is the SI unit of solid angle. Inverting this relation yields an exact expression for the candela: $1\ \mathrm{cd} = (K_{\mathrm{cd}}/683)\ \mathrm{kg\,m^{2}\,s^{-3}\,sr^{-1}}$ [48]. In this case, the ideal reference object is a perfectly monochromatic light source, specifically a laser or another light-emitting device that emits radiation at a frequency of 540×10^{12} Hz, which corresponds to green light. The property utilized is the luminous efficacy of this radiation, which describes the perceived brightness or intensity of the light per unit of energy. By fixing this value, the candela is defined as the luminous intensity of such a source emitting this frequency of radiation, providing a universal and reproducible standard for measuring light intensity.

4.8 Case Study: Definition and Realization of Meter

This section takes the Meter as an example to explain the definition and realization of the fundamental quantity.

4.8.1 Historical Definitions of Meter

Initial Definition of the Meter (1793) In 1793, the meter was defined as $1/10,000,000$ of the Earth's meridian quadrant [143]. *The Earth is the ideal reference object. Its property, the meridian quadrant, is utilized to define the unit of measurement.*

Before this definition, Europe used a variety of different units (such as feet, inches, and leagues), which differed widely across regions. To standardize measurements, the French Academy of Sciences proposed defining the meter as one ten-millionth of the distance from the North Pole to the Equator along the Paris meridian. This definition was a groundbreaking attempt to define a unit based on an unchanging natural property of the Earth, accessible to any nation through measurement. While subsequent definitions improved precision, the original concept, which rooted the meter in the Earth itself, remains central to the spirit of metrology.

To determine this length, astronomers Jean-Baptiste Joseph Delambre and Pierre Méchain undertook a monumental geodetic survey (1792-1799), measuring the meridian arc from Dunkirk to Barcelona [16]. This data formed the basis for calculating the full quadrant. During the survey, a provisional meter was created in 1793 based on the

preliminary results. However, the formal legal definition enacted that year was based on the survey's final results.

The International Prototype Meter (1875) In 1875, the CGPM (General Conference on Weights and Measures) redefined the meter using a platinum-iridium alloy bar (90% Pt, 10% Ir) [135]. This became known as the International Prototype Meter (IPM), or Bar No. 27, which was stored at BIPM (Bureau International des Poids et Mesures/International Bureau of Weights and Measures). *The ideal reference object is a platinum-iridium alloy bar (90% Pt, 10% Ir). Its property, the durability and low thermal expansion (8.7 µm/m°C), is utilized to define the unit of measurement.* The bar needed to be measured at 0°C. The design featured an X-shaped cross-section, with a total length of 102 cm, and two engraved lines marking 1 meter between their midpoints. Thirty copies were distributed to member nations for calibration, with Bar No. 6 being sent to the USA. Despite its revolutionary nature for 19th-century metrology, the physical standard posed risks, such as the potential for damage, microscopic wear, and accessibility challenges.

The Krypton-86 Definition (1960) The CGPM redefined the meter in 1960 using the emission spectrum of krypton-86 (^{86}Kr), marking the first definition based on a natural constant [41]. *The ideal reference object is krypton-86 (^{86}Kr), and its emission spectrum is utilized to define the unit of measurement.* The new definition specified that:

> 1 m = 1 650 763.73 wavelengths of the orange-red spectral line emitted
> by the ^{86}Kr isotope's transition between energy levels 2p and 5d,
> with 2p and 5d denoting specific atomic energy levels.

This definition achieved an accuracy of ±4 parts per billion (ppb), where 1 ppb denotes one part in a billion, i.e., a relative uncertainty of 10^{-9}, surpassing the limitations of the platinum-iridium bar. It used a discharge lamp filled with pure Kr vapor, excited electrically to emit the reference wavelength (605.780210 nm in vacuum). The new standard enabled global laboratories to independently realize the meter without needing to compare physical artifacts.

Transition to the Light-Speed Definition (1983) The 1960 krypton-86 definition demonstrated the SI system's adaptability to scientific progress. It laid the groundwork for the 1983 redefinition of the meter, based on the speed of light [201]. While the krypton-86 definition has since been superseded, it was a significant step in establishing metrological principles, particularly the use of atomic phenomena as the basis for measurements.

4.8.2 State-of-the-Art Definition of Meter

In 1983, the meter was defined as fixing the numerical value of the speed of light in vacuum (c) to be exactly 299 792 458 when expressed in the unit $\mathrm{m\,s^{-1}}$ [201]. *The ideal*

reference object is the light in vacuum, and its property, the constant speed, is utilized to define the unit of measurement. This definition indicates that: 1 m = c/299,792,458 s, where the speed of light in vacuum is defined as $c = 299{,}792{,}458\text{m s}^{-1}$ (exact).

4.8.3 State-of-the-Art Realization of the Meter

The current definition of the meter is based on fixing the speed of light in a vacuum, and its realizations must be traceable to atomic time standards [48]. Among the various implementation methods, one approach is officially recommended by the BIPM as a primary standard for realizing the meter: the iodine-stabilized helium-neon laser.

- *Principle*: Wavelength locked to $^{127}I_2$ transition R(127) at 632.991 nm.
- *Components*:
  - HeNe laser (633 nm).
  - Temperature-controlled iodine cell (±0.01 °C).
  - Fabry-Pérot cavity (finesse > 100).

The following outlines the replication requirements for primary standards, including conditions for vacuum, thermal control, vibration, and traceability.

- *Vacuum*: $\leq 1 \times 10^{-6}$ mbar for primary standards.
- *Thermal Control*: ±0.1 °C (lab), ±1 °C (industrial).
- *Vibration*: $< 10\,\text{nm}\,\text{s}^{-2}$ RMS.
- *Traceability*: all secondary measurement equipment must have valid calibration certificates traceable to national standards.

4.9 Summary

This chapter presents our interpretation of Metrology. Different from the classical definition of measurement: "objective obtaining of one or more values attributed to a quantity [17]", I define measurement based on comparing with a reference object, which reads "attributing values to a quantity of an object by comparing its quantity with the unit quantity of a reference object."

Also, I state the measurement problem and basic assumption, which serve as the axiom system for measurement. This thinking paradigm makes us ponder the essence of measurement and the fundamental role of measurement in interrogations. The historical reflections of the realization of fundamental quantities confirmed our thinking paradigm.

Chapter 5

Testology: The Science of Testing and Its Application

I propose to present the universal testing principles and methodology across different domains. I coined a new term, *Testology*, to cover this area. I define Testology as the science of testing and its application. Especially regarding the verification of theories and hypothesis testing, I believe these should fall under Testology, as they inherently follow the same testing principles. Dr. Lei Wang and I implemented this idea.

This chapter provides our interpretation of testing, including the basic concepts, problem statement, basic assumptions, fundamental principles, methodologies, and exemplary cases of testing.

I contributed to Sections 5.1, 5.2, 5.3, 5.4, and 5.5. Dr. Lei Wang contributed to Sections 5.6, 5.7, 5.8, 5.9, 5.10, and 5.11.

5.1 Basic Concepts

For better reading, we repeat some definitions in Section 3.6. A *proposition* is a testable statement about an object. A *model* is a streamlined representation of an object [221, 224]. A *fact or truth* is a proposition or a model about an object that can be proven true or verified objectively by a subject other than itself.

Test inputs refer to the data or stimuli provided to the object under test to drive the test. These may include parameters, user inputs, or configuration settings. For example, in a login system, the input might be a username and a password.

Test preconditions are the conditions set to ensure the object under test is in a valid state before running the test. Preconditions may involve system configurations or external conditions (such as power supply) that need to be met. For instance, for the same login system, a precondition could be that the user account must already exist in the database before testing the login functionality.

Test execution procedures define the steps or actions taken to run the test. In the

case of the login system, the execution procedure might involve opening the login page, entering the username and password, and clicking the "Login" button.

5.2 Definition of Testing

In this section, I present several essential testing concepts.

A test oracle is a ground truth or fact about an object. Additionally, for an artifact, a test oracle could be the intended behavior expected by the subject.

A test case is a predefined interrogation condition for testing, including test inputs, preconditions, and execution procedure, that is designed and implemented for a test oracle and is ready for executing an object under test to verify whether the actual outcomes are consistent with the mandated or expected results defined by the test oracle.

Testing is a verification process of running test cases to determine whether a proposition or a model of an object conforms to the test oracle through comparing their outcomes [224, 223]. *Being testable or testability* means a proposition or a model of an object can be falsifiable through testing.

5.3 Problem Statement

The testing problem could be stated as follows.

Given a test, how to design a set of test oracles and design and implement a corresponding set of test cases to balance the tradeoff between the testing accuracy and overhead?

The input of the testing problem is the object under test, while the output consists of a set of test oracles and a corresponding set of test cases that meet the stakeholders' requirements for testing accuracy and overhead.

5.4 Fundamental Assumptions

The most fundamental assumption in testing is that, for any object under test, we can always derive a complete set of test oracles and design and implement a corresponding complete set of test cases. Based on these, we can conclusively verify whether the object under test meets all testing criteria with 100% accuracy in the ideal scenario.

5.5 Fundamental Role of Testing

Testing is one of the fundamental interrogation methodologies, and its primary role is to falsify a proposition or a model about an object. Moreover, testing is the only means of interrogation that can validate a proposition or model as true or false.

From this perspective, it serves as a pillar for reasoning, which we will discuss in Chapter 6, as reasoning will generate different propositions or models, which can only be falsified by testing.

Meanwhile, it also serves as a pillar for evaluation, through which a subject could infer the effect induced by a cause object, which we will discuss in Part III. As inference is also reasoning, the inferred effect can also only be falsified by testing.

Just as the evaluation discussed in Section 2.2.9, testing is also practiced in an ad-hoc manner across different areas without a consensus on universal principles and methodologies. For example, testing is widely utilized in both hardware and software.

The fundamental role of testing is the reason why we coined a new term *Testology* to describe this area, the goal of which is to propose universal testing principles and methodologies.

5.6 Basic Principles

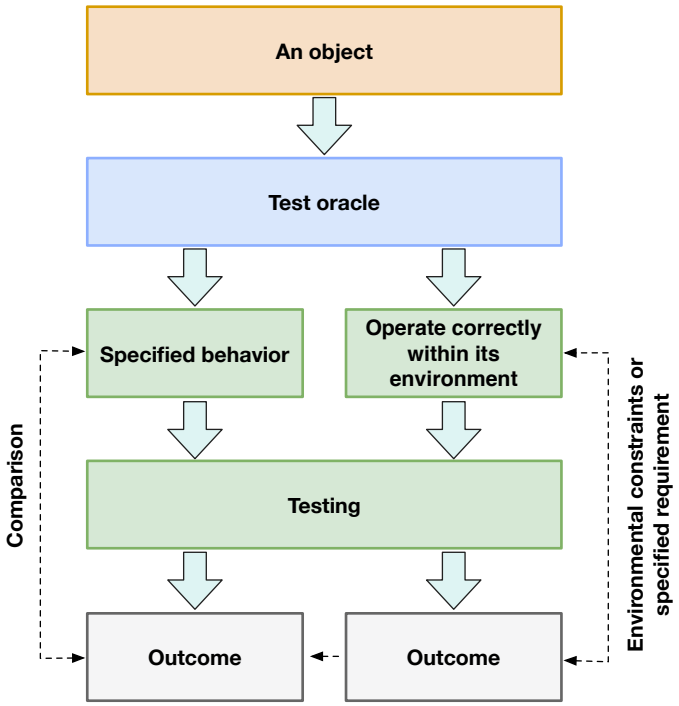


Figure 5.1: A simplified yet systematic conceptual framework for testing [13, 215].

As illustrated in Figure 5.1, the central concept of testing is the *test oracle*. A test

oracle functions as a reference, which we defined in Section 4.1, enabling the tester to distinguish between acceptable and unacceptable outcomes. The mandated or expected outcomes defined by a test oracle may be deterministic (exact values) or statistical (statistical values).

The process of testing involves running test cases to execute an object under test in predefined interrogation conditions defined by the test cases and subsequently comparing the outcomes against those mandated by the test oracles.

Two principal categories of test oracles may be distinguished. The first category focuses on verifying whether the object under test exhibits the outcome mandated by the test oracle. For instance, a sorting algorithm should return an outcome sequence in non-decreasing order for any valid input sequence.

The second category emphasizes the varying environments in which the object under test runs. The goal is to verify whether the object under test exhibits the outcome mandated by the test oracle in varying environments. This category of testing is to ensure that the artifact functions properly within the context under which it was designed, interacts appropriately with external services, complies with environmental constraints, and maintains stability under realistic operating conditions.

We further illustrate the principles of two typical testing methods: *Probability-based Testing* and *Deterministic Testing*, which differ in the test oracle.

5.6.1 Principles of Probability-based Testing Methodology

Probability-based Testing relies on a generalized ground truth: *under a hypothesis, if an observation outcome is unlikely, we would rather reject the hypothesis.*

In section 3.4, we have formally defined what a hypothesis is. A common example of probability-based testing is *hypothesis testing*, which tests whether an observed result is unlikely. The test oracle is: if the p-value, that is, the probability under the assumption that the null hypothesis is true, is less than the predetermined significance level α (e.g., 0.05), the null hypothesis H_0 is rejected.

For instance, in a clinical trial evaluating a new drug, the null hypothesis H_0 may assert that the drug does not affect blood pressure, implying that the average change in the Treatment Group is zero. The observed data are compared to this null hypothesis, and if the observed mean change significantly differs from zero (e.g., p-value < 0.05), the null hypothesis H_0 is rejected, suggesting that the drug likely affects blood pressure.

5.6.2 Principles of Deterministic Testing Methodology

Deterministic Testing verifies whether a system behaves exactly as expected under predefined conditions, producing a *definitive pass or fail* [215].

In deterministic testing, the test oracle is usually a set of predefined specifications or requirements, and correctness is determined by comparing whether the system's behavior matches these expectations.

For example, in the field of computing, two major forms of testing are *software testing* and *hardware testing*. Software and hardware are essentially artifacts, which we defined

in Section 3.3. Those methodologies could be extended to other artifacts.

- *Software testing* uses *test oracles*, such as specifications, reference implementations, or previous system versions, to determine correctness [37]. Failures can be classified as functional, when actual outcomes deviate from expected outcomes, or environmental, when constraints such as memory, performance, or compatibility are violated [13]. These failures could be cascading: for example, insufficient memory (an environmental failure) may cause the application to crash, resulting in a functional failure.

  A concrete example is testing a web application login: the test oracle specifies that valid credentials should return a successful login message. If the application crashes or returns an error despite correct credentials, it is a functional failure. If the server has insufficient memory, causing the application to slow down or become unresponsive under heavy load, it is an environmental failure that may indirectly trigger functional failures.

- A *test vector* consists of input sequences, expected outcomes, and fault models, simulating real-world operations to detect design or manufacturing defects. Test vectors are widely used test oracles in hardware testing for CPUs and digital circuits.

  For example, consider a 64-bit CPU with a 64-bit unsigned adder. A test vector may apply inputs A=0x0000000000000001 and B=0x0000000000000002. The expected outcome is S=0x0000000000000003 with carry-out=0. If the CPU produces the wrong sum or carry-out, it indicates a functional failure. Additionally, if a register has a stuck-at-0 fault (modeled in the fault model), the CPU will fail to produce the correct outcome, revealing a hardware defect.

5.7 Basic Methodologies

The basic methodology of testing is designing a subset of test oracle, designing and implementing a corresponding subset of test cases, and finally executing a subset of test cases to determine if the test passes by comparing the test results with those of the set of test oracle.

The testing process can be generalized into a unified, high-level framework, consisting of the following sequential steps:

1. *Test Oracle Definition*: A test oracle is defined to specify the mandated outcomes of a proposition or a model about an object.

2. *Test Case Design for a Test Oracle*: A set of test cases, including test inputs, preconditions, and execution procedures, is defined to compare the object's behavior against the expected outcomes described by the test oracle.

3. *Test Case Implementation*: According to the test case design, the test cases are implemented taking into account the characteristics of different environments.

4. *Test Case Execution*: The test case is executed, and the actual outputs or behaviors are then compared with the expected outcomes defined by the test oracle. Based on this comparison, a pass/fail determination is made for each test case.

In the following sections, we introduce the application of Testology in three areas: verifying a theory in Physics, significant testing in statistics, and software and hardware testing in computers. There are other applications of Testology, and we consider drafting a book on this topic in the future.

5.8 Verifying a Theory

We take the verification of parity non-conservation as an example to illustrate how to test a theory.

Parity (P) symmetry is one of the fundamental symmetries in classical physics and early quantum theory, describing the invariance of physical laws under spatial inversion [73]. Formally, a parity transformation replaces the spatial coordinates of a system (x, y, z) with $(-x, -y, -z)$, effectively converting a right-handed coordinate system into a left-handed one. A physical process is said to *conserve parity* if it remains unchanged under this transformation, meaning that its mirror-reflected version—expressed in the opposite-handed coordinate system—is physically indistinguishable from the original.

For a long time, many interactions, including electromagnetic and strong interactions, were believed to obey this symmetry. However, in 1956, Tsung-Dao Lee and Chen-Ning Yang conducted an in-depth analysis of experimental data and theoretical assumptions, proposing the hypothesis that parity might not be conserved in weak interactions [113]. They also put forward specific experimental ideas to detect potential asymmetries in particle emission directions. Parity non-conservation implies that the mirror image of a physical process—equivalently, its realization in the opposite-handed coordinate system—may not occur in nature, leading to observable asymmetries in particle behavior. In other words, under weak interactions, nature is capable of distinguishing between left-handed and right-handed coordinate systems.

In 1957, Chien-Shiung Wu and her team first directly verified parity non-conservation in weak interactions through experiments on cobalt-60 (Co^{60}) nuclei [219]. In the experiment, Co^{60} nuclei emit electrons via β-decay. The team used *adiabatic demagnetization* to polarize the Co^{60} nuclear spins at extremely low temperatures and employed a *single anthracene crystal scintillation detector*. By reversing the polarization magnetic field and observing changes in the counting rate of the same detector, they measured the asymmetry in the electron angular distribution. This approach effectively eliminated systematic errors caused by efficiency differences among multiple detectors. The experimental results clearly showed that electrons were emitted preferentially in the direction opposite to the nuclear spin, a phenomenon later confirmed through control experiments

to rule out possible magnetic artifacts. This discovery completely overturned the previous widespread belief that "parity is conserved in all interactions," exerting a profound impact on particle physics theory and experimental methods.

To systematically understand the scientific methodology of Wu's experiment, the *Testology* framework can be adopted.

In the definition of *test oracle*, Wu's testing experiments state that: in β-decay, if parity is conserved, the spatial distribution of emitted electrons along the nuclear spin axis should be symmetric. Any observed bias in the electron emission direction would indicate parity violation.

The *test case design* is a predefined experimental setup, including test inputs, preconditions, and execution procedures as follows:

- *Inputs*: High-purity Co^{60} radioactive sample (with known decay half-life and intensity).
- *Preconditions*:
    1. *Nuclear Polarization Apparatus*: The experiment used the Rose-Gorter method via adiabatic demagnetization of a cerium magnesium nitrate crystal to polarize Co^{60} nuclei.
    2. *Integrated Beta-Particle Detector*: A thin anthracene crystal scintillator was placed inside the vacuum chamber above the ^{60}Co source to detect the emitted beta particles.
    3. *Polarization Monitoring System*: Two additional NaI gamma-ray scintillation counters were installed—one in the equatorial plane and one near the polar position to monitor the polarization state of the sample.
    4. *Sample Preparation*: The Co^{60} sample was grown as a thin crystalline layer on the upper surface of good single crystals of cerium magnesium nitrate.
    5. *Control Experiments Preparation*: Two control specimens were prepared to eliminate potential artifacts from magnetic effects during the experiment.
- *Execution Procedures*:
    1. *Sample Mounting*: The Co^{60} sample crystal was mounted at the center of the demagnetization apparatus.
    2. *Polarization Process*: Adiabatic demagnetization was performed, followed by activation of the vertical solenoid to align the spins (20-second process).
    3. *Data Acquisition*: Beta and gamma counting commenced immediately:
        - Beta pulses were analyzed using a 10-channel pulse-height analyzer (1-minute intervals).
        - Gamma anisotropy was continuously monitored as an indicator of polarization.

4. *Field Reversal*: The polarizing field direction was reversed in subsequent runs to verify the authenticity of any observed asymmetry.
5. *Results*: The experimental data showed clear asymmetry effects, with electron emission significantly favored in the anti-spin direction.
6. *Control Verification*: Control experiments confirmed the absence of asymmetry under non-polarizing conditions.

In the *Test case implementation* stage, the experimental design is transformed into practical operations: the nuclei are magnetized at low temperature to orient their spins, the detectors are calibrated and arranged along the nuclear spin axis, and the instrument sensitivity and background noise are strictly controlled to ensure reliable data.

In the *Test case execution* stage, the electron counting rates along the nuclear spin direction and the opposite direction are recorded. If parity were conserved, the distributions in both directions would be symmetric. Wu's experiment, however, observed a significant bias in electron emission towards the anti-spin direction, providing direct evidence of parity non-conservation in weak interactions.

So, the universal methodology in *Testology* can be used to verify a theory.

5.9 Hypothesis Testing

Hypothesis testing is a statistical procedure that uses sample data to evaluate a null hypothesis (H_0) against an alternative (H_1) [132]. It employs a test statistic to measure the discrepancy between the data and H_0, and a decision rule (based on comparing the p-value to a significance level α) to either reject or fail to reject H_0.

In hypothesis testing, the *test oracle* comprises the null hypothesis (H_0), the alternative hypothesis (H_1), and the significance level (α), which together set the criteria for assessing whether the object's behavior significantly deviates from expectation. The *test case input* is the sample data, while the *test execution procedures* involve calculating the test statistic according to the *test oracle*. Finally, based on the calculated test statistic and the decision criteria, making the final decision to either reject H_0 or fail to reject H_0 is a pass/fail determination.

Under the *Testology* framework, hypothesis testing can be described as follows.

1. *Test Oracle Definition*: In hypothesis testing, the *test oracle* comprises the null hypothesis (H_0), the alternative hypothesis (H_1), and the significance level (α), which together set the criteria for assessing whether the object's behavior significantly deviates from expectation.

    - *Null Hypothesis (H_0)*: the observed outcome is consistent with the theoretical model, indicating no difference (e.g., the mean blood pressure of patients taking a new drug is equal to the population mean, $\mu = \mu_0$).

- *Alternative Hypothesis (H_1)*: the observed outcome show a statistically significant deviation from what is expected under H_0 (e.g., the mean blood pressure of patients taking the new drug is different from the population mean, $\mu \neq \mu_0$).

The decision rule depends on the significance level (α) and the p-value:

- *Significance Level (α)*: The pre-specified probability of rejecting H_0 when it is actually true. Commonly, $\alpha = 0.05$ is used:

$$P(\text{Reject } H_0 \mid H_0 \text{ true}) = \alpha. \tag{5.1}$$

- *p-value*: The probability of obtaining a test statistic at least as extreme as the observed one, assuming H_0 is true. The p-value computation depends on the type of test:
  - *Right-tailed test:* tests if the mean is greater than the hypothesized value ($H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0$).

$$p\text{-value} = P(X \geq x_{\text{obs}} \mid H_0). \tag{5.2}$$

  - *Left-tailed test:* tests if the mean is less than the hypothesized value ($H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0$).

$$p\text{-value} = P(X \leq x_{\text{obs}} \mid H_0). \tag{5.3}$$

  - *Two-tailed test:* tests if the mean is different from the hypothesized value ($H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$).

$$p\text{-value} = 2P(X \geq |x_{\text{obs}}| \mid H_0). \tag{5.4}$$

The decision rule is:

$$\text{Reject } H_0 \text{ if } p\text{-value} \leq \alpha.$$

2. *Test Case Design*: The test case is based on the sample data. The test statistic quantifies the discrepancy between the sample estimate and the null hypothesis. It is generally expressed as:

$$X = \frac{\hat{\theta} - \theta_0}{\text{SE}(\hat{\theta})}, \tag{5.5}$$

where $\hat{\theta}$ is the estimator derived from the sample, θ_0 is the parameter value under H_0, and $\text{SE}(\hat{\theta})$ denotes its standard error. The test case design specifies how to calculate the test statistic under the test oracle, including:

- *Test inputs (sample data)* to compute the estimator $\hat{\theta}$;
- *Preconditions*: primarily include assumptions about the underlying distribution, whether the variance is known or unknown, and the independence of samples.

- *Execution Procedures*:
  (a) *Calculate the test statistic:* Depending on the preconditions, and common test statistics include:
    – *t-test* (for population variance unknown):
    $$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}, \tag{5.6}$$
    where $\bar{X}$ is the sample mean, μ_0 is the population mean under the null hypothesis H_0, s is the sample standard deviation, and n is the sample size.
    – *z-test* (for population variance known):
    $$z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{5.7}$$
    where $\bar{X}$ is the sample mean, μ_0 is the population mean under the null hypothesis H_0, σ is the population standard deviation, and n is the sample size.
  (b) *Derive the p-value:* After obtaining the observed test statistic x_{obs} (t or z), the p-value is computed based on Equation 5.2, Equation 5.3, or Equation 5.4.

3. *Test Case Implementation*: This step transforms the test case design into an executable statistical procedure, for example, by looking up values in statistical tables or by writing a program in MATLAB to automate the calculations.

4. *Test Case Execution*: Execute the implemented test case and obtain actual outputs:
   (a) Calculate the observed test statistic x_{obs};
   (b) Calculate the corresponding p-value;
   (c) Apply the oracle decision rules:
   - If p-value $\leq \alpha$, *Reject* H_0;
   - If p-value $> \alpha$, *Fail to reject* H_0.

To illustrate this framework, consider a clinical trial conducted to evaluate the effect of a new drug on blood pressure. The null and alternative hypotheses are: $H_0 : \mu = 0$, the drug has no effect on blood pressure; $H_1 : \mu \neq 0$, the drug affects blood pressure. μ is the population mean change in blood pressure in the treatment group. The sample data if from 20 patients and obtain a sample mean and standard deviation of: $\bar{X} = 0.774$, $s = 2$. For population variance unknown, we use the t-test, the t statistic is calculated as: $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{0.774-0}{2/\sqrt{20}} = 1.73$. The two-tailed p-value is: $p\text{-value} = 2P(T \geq |t|) = 0.1$. Since p-value $= 0.1 > 0.05$, we fail to reject the null hypothesis. There is not enough statistical evidence to conclude that the drug significantly affects blood pressure.

In summary, hypothesis testing provides a probabilistic decision framework for assessing whether observed system outcomes are consistent with an expected model (H_0) or whether sufficient evidence exists to support an alternative hypothesis (H_1).

5.10 Software Testing

Software testing aims to verify whether a software system behaves as expected and satisfies its specified requirements. Failures detected during testing can be categorized into functional and environmental failures. Functional failures occur when the system produces incorrect results, while environmental failures—such as insufficient memory or slow execution—may indirectly lead to functional failures [13].

Under the *Testology* framework, software testing can be described as follows.

1. *Test Oracle Definition*: A test oracle is defined to specify the expected outcomes of a system or software component. The purpose is to ensure that the system's behavior aligns with the expected results. For instance, in a login system, if a user enters a valid username and the correct password, the expected outcome according to the test oracle is that the login is successful and the user is redirected to the homepage.

2. *Test Case Design*: A test case is defined, including specific inputs, preconditions, and execution procedures, to compare the system's behavior against the expected outcomes described by the test oracle. For example, in a login system, the input might be a username "admin" and the correct password; a precondition might be that the user account "admin" already exists in the system's database; and the execution procedure might involve opening the login page, entering the username and password, and clicking the "Login" button.

3. *Test Case Implementation*: According to the test case design, the test cases are implemented by creating the necessary test artifacts (e.g., test scripts, input files, or execution instructions), setting up the required testing environment, and ensuring that all inputs, preconditions, and execution steps are correctly instantiated. For instance, to implement the login test case, a tester could create a test script that opens the login page, inputs username "admin" and the correct password, clicks "Login" and verifies that the system redirects to the homepage. This ensures that the test case is executable and can reliably verify the expected behavior across different environments.

4. *Test Case Execution*: The implemented test cases are executed to verify whether the system behaves as expected. For example, the login test case would be executed by running the test script or performing the manual steps: open the login page, enter username "admin" and the correct password, click "Login" and observe the system's response. The actual output is compared against the expected outcome defined by the test oracle (redirection to the homepage). Based on this comparison, a determination is made that the test passes if the user is successfully redirected, or fails if the expected behavior does not occur.

Software testing methodologies can be categorized along different dimensions. Based on the level of knowledge about the system's internals, we have:

- Black-Box Testing. Black-box testing focuses on validating the functionality of the system without knowledge of its internal workings. This testing is based purely on the system's inputs and outcomes. For example, on an e-commerce site, "equivalence partitioning" can be used to test various inputs for a login form, such as ensuring that email addresses of different formats (valid or invalid) trigger the appropriate system response. "Boundary analysis" might be applied to test password length, ensuring that both the minimum and maximum character limits are correctly enforced by the system.

- White-Box Testing. White-box testing ensures that the system's internal logic is functioning correctly and aims to maximize test coverage. For example, on an e-commerce site, "code coverage" can be used to validate that every part of the code base—such as shipping calculations—is adequately tested. "Mutation testing" involves introducing small, controlled changes (mutations) to the code base and running tests to ensure that the test suite can detect these errors, ultimately confirming the thoroughness of the tests.

These approaches can be further implemented through different means:

- Manual Testing. The goal of manual testing is to ensure that the system functions as expected from the user's perspective. For example, consider an e-commerce site; testers may manually simulate the checkout experience to verify that all form fields are correctly populated and that users can successfully complete a purchase.

- Automated Testing. Automated testing utilizes scripts and tools to execute test cases without human intervention, enabling efficient and frequent regression validation. This approach ensures that previously verified features continue to function correctly after system updates, changes, or new feature deployments. For example, in an e-commerce site, automated tests can be programmed to simulate a wide range of user actions—such as user registration, product search, adding items to the shopping cart, and completing the checkout process—and automatically verify that the actual outcomes at each step match the expected results. The key advantage of automation is the ability to rapidly repeat these critical validation steps with consistent accuracy, significantly improving testing coverage and efficiency compared to manual execution.

5.11 Hardware Testing

The test vector approach is widely used in hardware testing for CPUs and digital circuits, and it involves predefined input sequences applied to a circuit (or hardware system) under test to verify its correctness against a test oracle [126].

Under the *Testology* framework, hardware test vector testing can be described as follows.

1. *Test Oracle Definition*: In hardware test vector testing, the test oracle is the reference used to determine the correct expected outcome for a given input vector V_i. It defines the expected behavior of the system under test by specifying the expected result $R_{\text{expected}}(V_i)$ for each input vector.

2. *Test Case Design*: A test case consists of input, preconditions, and execution procedures. A specific input vector V_i is the input. Test case preconditions include the necessary system state setup, such as powering on the device, initializing registers, and ensuring the circuit is in a known starting state. Execution procedures define the exact steps to apply V_i to the circuit and measure the output.

3. *Test Case Implementation*: Implementing a test case entails preparing the input vector and setting up the hardware test environment. Specifically, this involves loading the input vector into the test equipment, initializing the hardware under test, and verifying that all preconditions are met to ensure reliable application of the stimuli.

4. *Test Case Execution*: During execution, the input vector V_i is applied to the hardware, and the actual outcome $R_{\text{actual}}(V_i)$ is observed. The actual output is then compared against the expected output defined by the test oracle. The system passes the test for V_i if $R_{\text{actual}}(V_i) = R_{\text{expected}}(V_i)$; otherwise, a failure is reported.

Hardware testing methodologies for CPU validation can be categorized along different dimensions. Based on the level of knowledge about the system's internal microarchitecture, we have:

- Black-Box Testing. Black-box testing in CPU testing focuses on validating the behavior of the processor's instruction set architecture without any knowledge of its internal workings. This method tests the processor's ability to execute instructions as expected. "Equivalence partitioning" is used to validate the behavior of specific instructions, such as the "ADD" operation, ensuring that different input ranges are handled correctly. For example, boundary checks for overflow might be used to ensure that adding two large numbers does not cause incorrect results due to overflow. "Fuzzing illegal opcodes" is another technique used in black-box testing, where random or malformed instructions are fed into the CPU to identify any potential vulnerabilities or undefined behaviors.

- White-Box Testing. White-box testing focuses on verifying the internal logic of the CPU by ensuring that the processor's internal structure and pipeline handle all possible execution paths. One common technique is "path coverage," which ensures that all possible paths through the processor's pipeline (such as handling various exceptions) are tested. Another approach is "mutation testing," where small, controlled changes (mutations) are made to the processor's logic (e.g., flipping a comparator) to check if the existing test suite can detect these errors.

These approaches can be further implemented through different technical means:

- *Simulation-Based Testing.* This is the most common approach, dynamically verifying CPU functionality by running test programs in a simulation environment. For instance, engineers develop or generate extensive test vector programs that are executed in a simulator, with results compared against a golden reference model. This method's advantage lies in its flexibility to simulate complex scenarios and long execution sequences.

- *Formal Verification.* Formal verification is a static verification approach that uses mathematical methods to prove certain aspects of design correctness exhaustively, with-

out running the test system. For example, critical properties of the microarchitecture, such as data consistency in the pipeline, can be formally verified to hold under all possible instruction sequences. This method is particularly strong at uncovering subtle, deep design errors that are difficult to hit with random testing.

- *Hybrid Methods.* Complex CPU validation often employs hybrid techniques. *Concurrent hybrid verification*, for instance, combines simulation's flexibility with formal verification's exhaustiveness. While simulation runs, formal tools analyze reachable states from the current state and may automatically generate new test vectors to cover unexplored paths, thereby more systematically exercising various corner cases.

5.12 Summary

Testing can only show the presence of violating the ground truth or fact, but can not prove that there is no existence of violating the ground truth or fact. Exhaustive Testing is Impossible. Testing identifies defects; debugging (by developers) fixes them. These are separate but complementary processes. For example, in software testing, testing can demonstrate that defects exist in the software, but it cannot prove that the software is defect-free. Even if no bugs are found, it does not mean the system is perfect—only that no defects were detected under the tested conditions. It is impractical to test every possible input, combination, or scenario, especially in complex systems. Instead, testers use risk analysis and prioritization to focus on the most critical areas.

Chapter 6

Reasoning

This chapter provides the basic concepts, problem statement, basic assumptions, fundamental principles, methodologies, and exemplary cases of reasoning. Dr. Fanda Fan and I co-authored this chapter. Dr. Fanda Fan authored Sections 6.6, 6.7, 6.8, 6.9, and 6.10. I authored Sections 6.1, 6.2, 6.3, 6.4, and 6.5.

6.1 Basic Concepts

For better reading, I repeat several definitions in Sections 3.4, 3.5, and 3.6.

A *proposition* is a testable statement about an object. A *model* is a streamlined representation of an object [221, 224]. A *fact or truth* is a proposition or model about an object that can be proven true or verified objectively by a subject other than itself. *Knowledge* contains facts or truths about objects. *An Evidence* is data about an object that supports, refutes, or informs a proposition or model.

A *premise* is a proposition or model that serves as the foundational statement from which a conclusion is logically derived. A *conclusion* is the proposition or model obtained through a reasoning process.

Hypothesis is a kind of proposition or model of an object. The *assumption* is a proposition or model accepted as truth or fact without proof, and it is a *provisional premise* adopted within a specific logical argument. *An axiom system* is *a collection of self-contained assumptions* for domain-specific knowledge or those beyond a specific domain.

6.2 Definition of Reasoning

I defined *reasoning or inference* as *a capacity, capability, and mental process of generating new propositions or models about an object based on the premises, the facts or truths, and the interrogation outcomes.* In the rest of this article, we do not distinguish between reasoning and inference if not explicitly stated.

A reasoning or inference rule is a rule that generates a new proposition or model based on the premises, the facts or truths, and the interrogation outcomes.

Prediction is a kind of reasoning to infer the interrogation outcome of an object according to its truth, fact, assumption, or model.

6.3 Problem Statement

Depending on different purposes, the reasoning problem can be stated in different ways. For those interested in the knowledge system, the reasoning problem can be stated as "how to propose a simple and concise axiom system, based on which we can build the corresponding knowledge systems."

For those who are interested in reasoning or inference itself, they care about the effectiveness or efficiency of reasoning. For the former purpose, the focus is on the outcome: the problem could be stated as "whether and how the desired propositions or models can be achieved, which is often referred to as proof." For the latter purpose, the focus is on the reasoning process: the problem could be stated as "how to reason according to the correct rules with minimal waste in terms of time, resources, and effort."

6.4 Fundamental Assumptions

There are two fundamental assumptions in reasoning.

First, it is assumed that several propositions are correct without the need for proof, which is often called the axiom system.

Second, there is a concise set of perfect reasoning rules, which can pass testing 100%. However, if a contradiction happens by testing, the reasoning rule is invalidated.

6.5 Fundamental Role of Reasoning

Reasoning is a mental capacity and capability, and only a subject can own this capability and capacity. A subject can use a mental activity to replace a real activity in the physical world.

From this perspective, reasoning has three fundamental roles. First, it augments the subject's capability to understand objects. The subject could generate new propositions or models about objects; Reasoning, as a mental activity, could replace physical activities in the real world, and relieve some tasks through reasoning.

Second, through reasoning, we can build a simple and concise axiom system for the knowledge system. For example, we can restructure or improve the measurement and testing methodologies by building them upon a different axiom system.

Third, through studying the nature of reasoning, we can understand the inherent flaws of our knowledge system, e.g., Gödel's Incompleteness Theorems [62] discussed in Section 3.6.4.

6.6 Basic Principles

Logical reasoning is grounded in *formal systems*: structured collections of symbols, rules, and axioms that specify how valid conclusions can be derived. Among these systems, *first-order logic (FOL)* [64] provides the canonical foundation. It consists of:

- **Variables** $x, y, z, \ldots$
- **Predicates** $P, Q, \ldots$
- **Logical connectives** $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$
- **Quantifiers** $\forall, \exists$

A typical expression takes the form $\forall x\,(P(x) \rightarrow Q(x))$, which asserts a universally quantified implication.

Operator	Symbol	Meaning
Negation	$\neg P$	Not P
Conjunction	$P \wedge Q$	P and Q
Disjunction	$P \vee Q$	P or Q
Implication	$P \rightarrow Q$	If P, then Q
Biconditional	$P \leftrightarrow Q$	P iff Q
Universal Quantifier	$\forall x$	For all x
Existential Quantifier	$\exists x$	There exists x

Table 6.1: Common logical operators and their meanings.

Axiom Systems in Logic

Based on these symbols and rules of inference, various *axiom systems* have been proposed to formalize valid reasoning. Classical logic begins with foundational principles such as:

Non-Contradiction (Aristotle). Aristotle's metaphysics [7] formulates the basic law:

$$\neg(P \wedge \neg P), \tag{6.1}$$

stating that no proposition P can be true and false at the same time. This principle underlies all coherent evaluative reasoning: contradictory claims cannot simultaneously support a valid conclusion.

Limits of Axiomatic Systems (Gödel). Kurt Gödel's incompleteness theorems [69] reveal a fundamental constraint: no sufficiently expressive axiomatic system can be both *complete* and *consistent*. That is, some true statements cannot be derived from the system's own axioms.

These results clarify an important point for reasoning in evaluative contexts: even with well-defined rules, *no formal system can capture all truths solely through axioms*. Thus, logical reasoning provides structure and rigor, but it does not eliminate the need for empirical testing, contextual analysis, or interpretive judgment.

First-order logic [64] is a formal system that includes individual variables $x, y, z, \cdots$, predicates $P, Q, \cdots$, quantifiers $\forall, \exists$, and connectives $\rightarrow, \leftrightarrow$, and takes forms like $\forall x(P(x) \rightarrow Q(x))$.

The reasoning is built upon the axiom system regarding logic. The Propositional and Predicate Logics are as follows (See in Table 6.1).

Based on the above logic, several axiom systems have been proposed. For example, Aristotle's ontological law [7] stated a Non-Contradiction Principle:

$$\neg(P \wedge \neg P). \tag{6.2}$$

No proposition P can be simultaneously true and false.

Kurt Gödel's incompleteness theorems [69] fundamentally demonstrate inherent limitations in formal axiomatic systems –sets of rules and symbols used to derive mathematical truths.

6.6.1 A Brief about Kurt Gödel's Incompleteness Theorems

First Incompleteness Theorem: Any consistent formal system powerful enough to express basic arithmetic (like Peano arithmetic) is *incomplete*. This means there will always be true statements expressible within the system that cannot be proven true using the axioms and rules of that system itself. The system contains "gaps" in provability.

Second Incompleteness Theorem: Such a system cannot prove its own consistency using only its own axioms and rules. If it's consistent, the statement asserting "this system is consistent" remains unprovable within the system.

6.6.2 How They Limit Formal Systems

In essence, Gödel [69] proved that any formal system rich enough to handle elementary arithmetic is either incomplete (missing proofs for some truths) or inconsistent (capable of proving contradictions). Furthermore, it can never certify its own soundness. This imposes fundamental, unsurpassable barriers on what any single formal system can achieve in capturing mathematical reality.

6.7 The Summary of Fundamental Reasoning Methodologies

There are three fundamental reasoning methodologies: *deductive reasoning*, *inductive reasoning*, and *abductive reasoning*. Each embodies a different way of moving from what is already known or observed to what is newly inferred.

Deductive reasoning proceeds from the general to the particular. Starting from accepted premises and axioms, it derives conclusions that follow with *logical necessity*: if the premises are true and the inference rules are correctly applied, the conclusion cannot be false. Deduction thus preserves truth by logical form alone, independently of empirical content.

Inductive reasoning proceeds from the particular to the general. It abstracts new propositions or models from finite collections of interrogation outcomes, yielding conclusions that are *probable* rather than certain. Induction is ampliative: the conclusion goes beyond the premises, supporting generalizations that may later be refined or overturned as new evidence arrives.

Abductive reasoning proceeds from interrogation outcomes to *plausible hypothesis*. Abduction does not guarantee truth, but proposes the *best available* hypothesis under current knowledge.

Each of the following sections adheres to a consistent reasoning structure. For every reasoning type, we begin with its *definitional principles*, clarifying its logical foundations and epistemic status; then provide an example to demonstrate its concrete process; follow with *methodological paradigms* that trace its historical and contemporary formalizations.

6.8 Deductive Reasoning

6.8.1 Definitional Principles

Building on the definitions of propositions, premises, and axiom systems introduced earlier, *deductive reasoning* provides the most stringent mechanism for generating new propositions. It proceeds from the general to the particular—deriving specific conclusions from universal premises with logical necessity. Unlike induction, which abstracts generalities from empirical observations, deduction operates entirely within a formal structure of axioms and inference rules. Once the premises are accepted as valid elements of the knowledge system, the resulting conclusion follows with certainty. In this sense, deduction establishes the *architecture of reasoning certainty*: truth is preserved by logical structure rather than experiential variability.

Formally, let $P_1, P_2, \ldots, P_n$ denote a set of *premises* (propositions or models), and let C denote a *conclusion*. The notation

$$P_1, P_2, \ldots, P_n \vdash_{\text{ded}} C, \tag{6.3}$$

indicates that, under the rules of deductive inference, C follows necessarily from the premises $\{P_i\}$. Here, $\vdash_{\text{ded}}$ denotes the *deductive entailment relation*, meaning that the

validity of the inference depends exclusively on logical form rather than empirical content.

A canonical example is the rule of *modus ponens*. Given a conditional proposition $P \to Q$ [1] and the premise P, deductive entailment yields:

$$(P \to Q),\ P \vdash_{\text{ded}} Q. \tag{6.4}$$

This inferential pattern remains valid regardless of the semantic content of P and Q—its justificatory force derives solely from logical structure.

Deductive systems rest on a small set of axioms and inference rules that generate an infinite set of valid propositions. The most common logical connectives are summarized in Table 6.1. Each connective defines a lawful transformation within propositional or predicate logic. Deductive reasoning, then, is the disciplined application of these transformations to derive conclusions already implicit in the premises.

A typical example of deductive reasoning

Premise 1: All even numbers are divisible by 2.
Premise 2: 14 is an even number.
Deductive conclusion: Therefore, 14 is divisible by 2.

6.8.2 Methodological Paradigms

Deductive reasoning has evolved from classical syllogistic forms to symbolic and computational logic, which now underlie mathematics and artificial intelligence.

Syllogistic Deduction. The earliest deductive framework traces back to *Organon* [6]. A syllogism takes the canonical form:

$$\text{All } A \text{ are } B;\quad \text{C is } A;\quad \therefore \text{C is } B. \tag{6.5}$$

Here, A, B, and C denote *classes of objects*. The statement "All A are B" asserts that every object belonging to class A also belongs to class B, while "C is A" asserts that the object C is a member of class A. Given these premises, the conclusion that C is also a member of class B follows with logical necessity.

This form illustrates that the conclusion introduces no new information beyond what is already contained in the premises. Classical syllogistic inference thereby established the principle that deductive validity depends solely on logical form. This principle later becomes crucial in determining whether propositions and models about objects remain structurally coherent under different interrogation conditions.

Formal and Mathematical Deduction. In modern logic, formal deduction was codified through the axiomatic systems of symbolic logic [64, 82, 165]. These frameworks define rules of inference independent of content or interpretation. For instance:

$$\forall x\ (A(x) \to B(x)),\ A(a) \vdash B(a). \tag{6.6}$$

[1] $P \to Q$: read as "if P, then Q"

This symbolic precision supports theorem proving and verification across mathematical and computational domains.

Computational Deduction. With the rise of computing, deduction became mechanized through automated reasoning frameworks, including logic programming [105] (e.g., Prolog), resolution-based theorem provers [9, 10], SAT/SMT solvers [8, 70], description logic reasoners [204], and model checkers [63]. These systems perform large-scale consistency checking, proof search, constraint satisfaction, and validation of complex rule sets. In artificial intelligence, computational deduction enables knowledge-base maintenance [211], formal verification of learning systems [50, 147], and the enforcement of safety constraints [112].

6.9 Inductive Reasoning

6.9.1 Definitional Principles

Inductive reasoning allows us to generalize a proposition or model from specific cases (like seeing multiple white swans and generalizing that all swans are white). However, this reasoning doesn't guarantee certainty, as new observations might contradict the generalization. The conclusion is probable, but not logically necessary.

This is different from deductive reasoning, which guarantees that the conclusion is true if the premises are true (e.g., if "all swans are white" and "this is a swan," then the swan must be white). In inductive reasoning, conclusions emerge based on a similarity across multiple interrogations, even though each individual observation might not guarantee the outcome.

Formally, we may write:

$$P_1, P_2, \ldots, P_n \vdash_{\text{ind}} C \quad \text{where } \forall i, P_i \nvdash_{\text{ded}} C. \tag{6.7}$$

That is, the conclusion C does not follow with logical necessity from any single premise P_i, but emerges as a plausible proposition or model across them all.

Three principles underlie valid induction:

1. **Evidence Accumulation.** Confidence in a *hypothesis* H grows as more consistent *evidence* $E_1, E_2, \ldots, E_n$ is observed. Under classical i.i.d. (independent and identically distributed) assumptions, repeated confirmation increases the posterior probability of H:

$$\lim_{n\to\infty} P(H \mid E_1, \ldots, E_n) = 1. \tag{6.8}$$

   Thus, reliable induction depends not on any single evidence item E_i, but on the convergence of evidence across repeated observations.

2. **Projective Invariance.** Inductive generalizations must remain stable when applied to different contexts or domains. Let $P(H)$ denote the prior probability of hypothesis H, and $P(H \mid E)$ its posterior probability after observing evidence E. The ratio
$$\frac{P(H \mid E)}{P(\neg H \mid E)} > \frac{P(H)}{P(\neg H)}, \tag{6.9}$$
expresses that evidence E should *increase* the odds in favor of H. A generalization worth preserving in one domain should, when properly abstracted, retain its inductive support across others.

3. **Error Bounding.** Every inductive conclusion involves uncertainty, which must be formally quantified. Let $\hat{\theta}$ denote a sample's estimator for the parameter of a population, $z_{\alpha/2}$ the critical value of a standard normal distribution at significance level α, and n the sample size. A classical α-level confidence interval CI_α takes the form:
$$\mathrm{CI}_\alpha = \hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}. \tag{6.10}$$
Inductive reasoning succeeds not by eliminating error, but by measuring, bounding, and managing it systematically.

A simple example of inductive reasoning

Observation 1: The sun rose in the east yesterday.
Observation 2: The sun rose in the east today.

Inductive conclusion: The sun *will* rise in the east tomorrow.

This conclusion does not follow with logical necessity from any single measurement or testing; Rather, it emerges from recognizing a consistent similarity across multiple observations. Hence, the inference is **probable** rather than **certain**, illustrating the nature of inductive reasoning.

6.9.2 Methodological Paradigms

Inductive reasoning is central to many methodologies, and it manifests across a variety of frameworks, each formalizing the logic of generalization under different epistemic assumptions. These frameworks provide tools for reasoning from observations, updating and fitting models, and making predictions according to a truth, fact, or hypothesis, each with unique strengths and applications. Below, we explore three key paradigms of inductive reasoning: the Bayesian framework, Statistical Learning Theory, and the Algorithmic Probability framework.

Bayesian Framework. The Bayesian framework [15, 92, 127, 86] dynamically updates hypotheses as new evidence arrives. This approach provides a principled mechanism for

updating hypotheses H based on observed evidence E, using the following expression:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}. \tag{6.11}$$

As evidence E accumulates, the posterior probability $P(H|E)$ strengthens or weakens accordingly. This framework is foundational to modern inductive reasoning.

Statistical Learning Theory. Within the Statistical Learning Theory framework [209], induction is framed as the process of model fitting and prediction. It formalizes the learning process as the search for a hypothesis function $\hat{f}$ that minimizes the expected loss:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda R(f). \tag{6.12}$$

Here, y_i represents the true output (or target) value for the i-th observation outcome, while x_i represents the corresponding input (or feature). The term L is the empirical loss function that quantifies the difference between predicted and true values, R is a regularization term that controls the complexity of the model, and λ is the trade-off parameter that balances model fit and generalization.

In this context, learning is an inductive process where the goal is to find the simplest model that best fits the observed outcomes while avoiding overfitting. The learner seeks a function $\hat{f}$ that generalizes well to unseen outcomes, ensuring that the model not only fits the observed outcomes, which is often called *training data*, well, but also maintains simplicity, preventing overfitting.

Algorithmic Probability Framework. At the theoretical frontier lies the Algorithmic Probability framework [180, 88], which defines the idealized limit of all computable inductive systems. This framework assigns a universal prior probability $P_U(x)$ to a data sequence x by summing over all programs p that produce x (or any extension of it) on a universal Turing machine U:

$$P_U(x) = \sum_{p:U(p)=x*} 2^{-\ell(p)}, \tag{6.13}$$

where $\ell(p)$ is the length of program p, and $U(p) = x*$ denotes that the output of p on U has x as its prefix.[2] Shorter programs—simpler explanations—receive exponentially higher prior probability, providing a formal realization of Occam's Razor [148]. Although incomputable in practice, this framework serves as a conceptual ideal that unifies probability, simplicity, and prediction into a single theory of rational inference.

Together, these paradigms reveal the gradient of inductive reasoning—from pragmatic updating (Bayesian) to statistical optimization (Learning) to universal reasoning

[2]Formally, $x*$ represents the set of all strings whose initial segment equals x. Thus, any program whose output begins with x contributes to $P_U(x)$, reflecting the framework's role in sequence prediction.

(Algorithmic). Each paradigm contributes a distinct lens through which to reason from evidence, with applications ranging from hypothesis updating to model fitting and prediction.

6.10 Abductive Reasoning

6.10.1 Definitional Principles

While deduction moves from general rules to necessary conclusions, and induction moves from repeated observations to probable generalizations, *abductive reasoning* proceeds in the opposite direction: it infers the *most plausible* hypothesis that could explain an observed outcome.

First introduced by Peirce—who characterized it as "the logic of discovery" [142]—abduction is the process of forming explanatory hypotheses. It begins with an observation and seeks an explanatory hypothesis. Abduction thus provides the conceptual bridge between observation and explanation. It does not ask "What must be true?" as in deduction, nor "What is probably true?" as in induction, but rather "What would best explain what we observe?"

Formally, let E denote an observed *evidence* and let H denote a candidate *hypothesis*. Abductive reasoning can be expressed as:

$$H \to E, \quad E \text{ observed} \quad \therefore \quad H \text{ (as a plausible hypothesis).} \tag{6.14}$$

This inference is inherently **non-monotonic**: the acceptance of H as a plausible hypothesis can be overturned by new evidence, additional hypotheses, or alternative hypotheses. Unlike deductive entailment, the relation

$$H \vdash_{\text{abd}} E, \tag{6.15}$$

does *not* guarantee that H is true—only that H is a *reasonable* explanation under the current knowledge.

A simple example of abductive reasoning

Observation: The ground is wet this morning.
Possible explanation H_1: It rained last night.
Possible explanation H_2: The sprinkler system ran overnight.
Abductive conclusion: The most plausible explanation is that it rained.

This conclusion may later be revised if new evidence appears (e.g., sprinkler logs). Abduction thus illustrates reasoning that is **explanatory but fallible**.

6.10.2 Methodological Paradigms

Abductive reasoning plays a foundational role in scientific modeling, diagnosis, hypothesis formation, and mechanism discovery. Its methodological paradigms include:

Inference to the Best Explanation. Inference to the Best Explanation [115] formalizes abduction as selecting the hypothesis that best explains the observed evidence, based on certain explanatory virtues. These virtues include criteria such as simplicity (the hypothesis is not overly complex), coherence (the hypothesis fits well with established knowledge), breadth (the hypothesis explains a wide range of phenomena), and explanatory power (the hypothesis accounts for the evidence in a compelling way).

Given multiple candidate hypotheses $\{H_1, H_2, \dots\}$, each of which could explain the evidence E, the abductive task is to select the one that maximizes an explanatory score:

$$H^* = \arg\max_{H_i} \text{Explains}(H_i, E), \tag{6.16}$$

where $\text{Explains}(H_i, E)$ quantifies how well hypothesis H_i explains E. This framework is qualitative, relying on reasoning about the explanatory virtues of each hypothesis.

Probabilistic Abduction. In contrast, probabilistic abduction takes a quantitative approach, viewing abduction through the lens of Bayesian inference. Here, abduction corresponds to selecting the hypothesis H that maximizes its posterior probability, given the observed evidence E:

$$H^* = \arg\max_{H} P(H \mid E), \tag{6.17}$$

where $P(H \mid E)$ is the posterior probability of the hypothesis H after observing evidence E. This approach provides a formal way of evaluating hypotheses, integrating prior knowledge about the hypotheses with the likelihood of the evidence under each hypothesis. Probabilistic abduction thus makes the reasoning process more rigorous and quantifiable by combining both prior knowledge and empirical data.

Model-based Abduction. Widely used in diagnosis and scientific modeling, this approach reasons over structured models (causal graphs [144], knowledge bases [121], or generative mechanisms [19]). The hypothesis H is considered plausible if incorporating it into the model makes the observed evidence E highly likely.

Together, these paradigms reflect the nature of abduction as a search for explanatory adequacy—a reasoning process that is creative, defeasible, and indispensable for forming new hypotheses.

6.11 Summary

Built upon axioms and inference rules, reasoning allows subjects to derive new propositions or models from existing truths, revealing the internal logic that governs both

knowledge systems. Deductive reasoning ensures internal validity by deriving necessary conclusions from established premises; inductive reasoning generalizes from repeated evidence to formulate empirical laws; and abductive reasoning infers the most plausible causes behind observed effects, driving discovery and interpretation. Together, these three modes form a continuous epistemic cycle—abduction proposes, induction confirms, and deduction secures coherence.

Chapter 7

Interrelationships Among Three Interrogations

In this chapter, I first posit the existence of a primitive interrogation among the three fundamental interrogations, and then proceed to examine the relationships between these three fundamental interrogations.

7.1 Primitive Interrogation: Comparison with a Reference

In Chapters 4, 5, and 6, I define measurement, testing, and reasoning as follows, respectively.

Measurement is a capacity, capability, and process that attributes values to a quantity of an object by comparing its quantity with the unit quantity of a reference object under an interrogation condition.

Testing is a capacity, capability, and verification process of running test cases to determine whether a proposition or a model of an object conforms to the test oracle through comparing their outcomes [224, 223].

Reasoning or inference is a capacity, capability, and mental process that generates new propositions or models about an object based on the premises, the facts or truths, and the interrogation outcomes.

Among the three fundamental interrogations, we could find a primitive interrogation, which is the comparison of an object with a reference object. Comparison is employed to establish the order among the same quantities of different objects. For measurement, the subject compares a quantity of any object with a unit of measurement of a reference object. For testing, the subject runs test cases and compares the outcomes with those of a test oracle. Hereby, a test oracle defines a reference. Reasoning relies upon reasoning rules. These rules can only be validated through testing, which requires a reference (test oracle) for comparison.

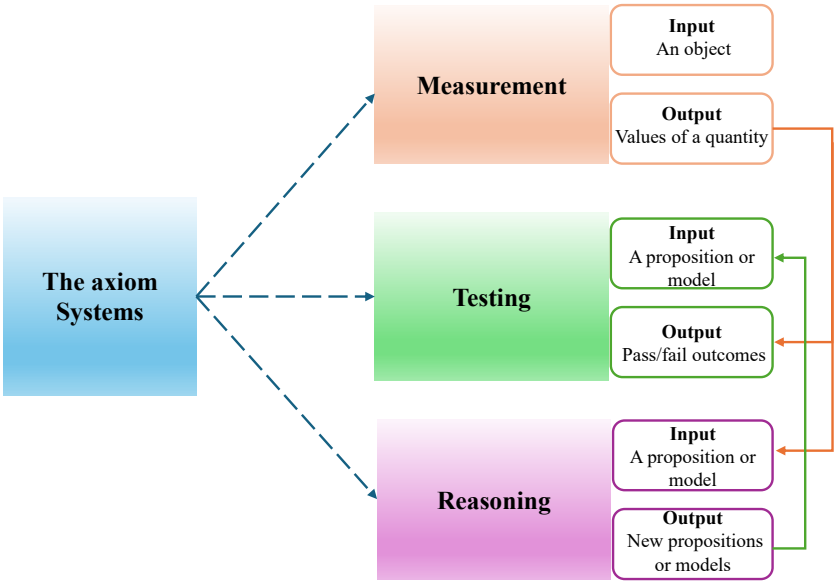


Figure 7.1: The relationship among three interrogations.

7.2 Relationships among Three Interrogations

The relationships among the three interrogations are shown in Figure 7.1 [1].

When testing is performed on a hypothesis about an object, it is essential to compare the outcome of running a test case to the test oracle. Most comparisons involve measured quantity values, which are inherently derived from measurement. That is to say, testing often relies on quantitative data obtained through measurement. The input for reasoning comprises propositions and models of objects, which can only be derived from quantitative data obtained through measurement. However, I believe many properties are not countable.

Measurement, testing, and reasoning rely upon an axiom system, which is one of the central topics in reasoning. For example, measurement relies upon an assumption that a property of an ideal reference object could be utilized to define the unit of measurement. However, the axiom system in measurement can only be validated by testing. Over the course of metrological evolution, numerous such ideal reference objects have been adopted and subsequently rendered obsolete following rigorous testing. Meanwhile, the inputs, outputs, intermediates, and even the reasoning rules in reasoning can only be validated through testing.

[1] Dr. Lei Wang contributed this figure based on the discussion with me.

7.3 Summary

This chapter presents that three fundamental interrogations are built on a primitive interrogation: comparison. I also clarify the interrelationships among measurement, testing, and reasoning.

Part III

The Science of Evaluation

Chapter 8

Basic Evaluation Concepts and Problem Statements

In this chapter, I present the basic evaluation concepts and problem statements.

8.1 Basic Concepts

For objects A and B, when measurable or testable differences occur in B depending on the presence or absence of A, I define A as the *cause object (in short, cause)*, B as the affected object (in short, AO), and the measurable or testable difference in B as the *effect* on B induced by A. Additionally, I define the *effect mechanism* as the way through which the cause object induces the effect on the AO. When I mention *mutual influences between objects A and B*, I refer to both A's effect on B and B's effect on A. Figure 8.1 shows the relationships among the cause, AO, effect, and effect mechanism.

Confounding occurs when two independent objects are associated in a manner that makes it challenging to differentiate their specific effects on a third object. In other words, the effects of these independent objects become entangled, making it difficult to attribute specific effects to each one.

In statistics, a *correlation* measures the mutual effect of variables a and b, among which many confounding variables exist. Correlation does not imply a cause-and-effect relationship.

Several Examples on Cause, AO, Effect, and Effect Mechanism.

An Physics Example

Issac Newton's Apple Tree: The *cause* is the earth, the *AO* is the apple, the *effect* is the gravitation, the *effect mechanism* is the Law of Universal Gravitation.

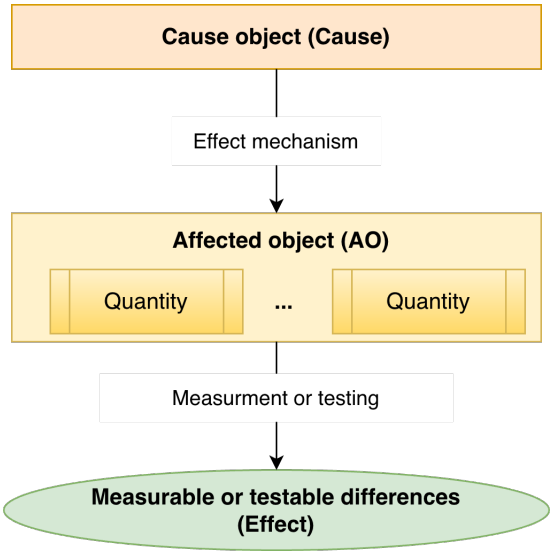


Figure 8.1: The relationships among Cause, AO, Effect, and Effect mechanism.

A Chemistry Example

Lavoisier's Combustion Experiment: The *cause* is oxygen, the *AO* is any substance that can be burned or calcined, the *effect* is combustion, the *effect mechanism* is the chemical reaction equation for combustion.

A Biology Example

Gregor Johann Mendel's Peas: The *cause* is the gene, the *AO* is the pea, the *effect* is the trait of the pea, and the *effect mechanisms* are the laws of segregation and independent assortment.

A Computer Science Example

CPU evaluation: The *cause* is the CPU, the *AO* is the computer system, the *Effect* is the CPU's effect on the measured overall performance metrics, and the *effect mechanism* is the collaboration mechanisms among the CPU and other indispensable components.

An Artificial Intelligence Example

Evaluating an AI algorithm: The *cause* is the AI algorithm, the *AO* includes the dataset labeled with ground truth, the algorithm implementation, the OS, the processor, and the memory. The *effect* is the AI algorithm's effect on the overall algorithmic accuracy, and the *effect mechanism* is the capability of an algorithm to fit input-output relationships.

A Medicine Example

Drug evaluation: The *cause* is the drug, the *AO* is the patients, the *Effect* is the measurable or testable difference in patients, and the effect mechanism is some known or unknown biological mechanism.

A Social Science Example

Policy evaluation: The *cause* is the policy, the *AO* is different participants, the *effect* is the policy effect on overall evaluation outcomes, such as the approval rating, and the *effect mechanism* is embodied in many forms, like political, social, or psychological ones.

8.2 Essence and Problem Statement

Based on the fundamental concepts presented in Chapter 8.1, I will define the concept of evaluation in this section and formally state the problem of evaluation.

When evaluating an object A, we need to identify the set of other objects B_i on which differences can be measured or tested, depending on the presence or absence of A.

We call A, the *evaluated object (in short, EO)* or the *cause object (in short, cause)*, B_i, the AO, and the measurable or testable difference in B_i, the *effect* on B_i induced by A. Essentially, the problem of evaluation can be formally stated as *how to uncover the effect of the EO on the AO*. If the instantiations of an object do not induce different effects, we do not distinguish between an object and its instance.

In addition to the EO A, other objects, A_j, also affect the AO, B_i, and we call those objects, the *essential external objects (in short, EXO)*. We call the effect induced by both the EO and the EXO on the AO *the overall effect*.

An EO exerts effects on numerous other objects. We categorize two distinct types: *a direct AO* and *an indirect AO*. A direct AO is *a minimal system on which the effect of the EO can be directly measured or tested*. An AO is either a direct AO or an indirect AO.

I formally call *the effect of the EO on the direct AO*, *the derived EO*, and call *the effect of the derived EO on the indirect AO*, *the derived effect*. In this context, the essence of evaluation is *how to uncover the effect of the EO on the direct AO and its derived*

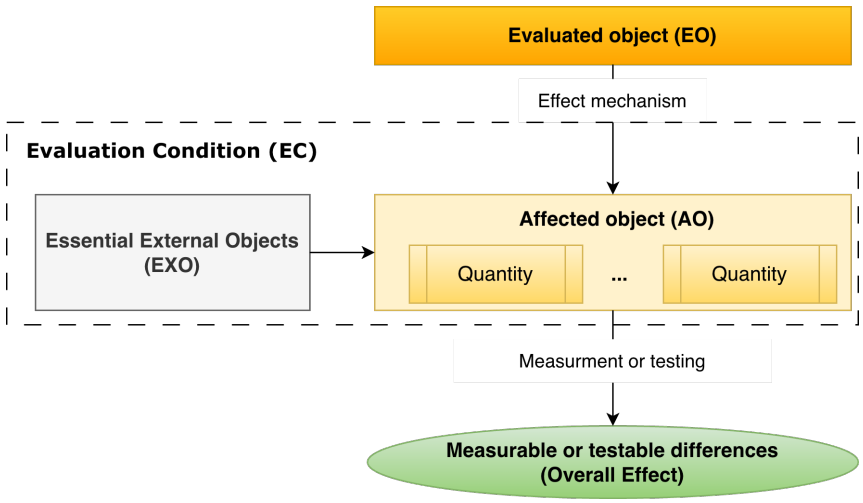


Figure 8.2: The fundamental components of an SES.

effect.

Some Examples on Direct AO and Indirect AO.

A Computer Example

A group of fans aims to snatch up concert tickets for their favorite superstar using computer devices equipped with different CPUs. An evaluation of the CPUs in this scenario is required. Here, the CPU serves as the EO, the computer system running the ticket-snatching program acts as the Direct AO, and the fans attempting to purchase the tickets function as the indirect AO.

A Physics Example

Issac Newton's Apple Tree: The EO is the Earth, the direct AO is the apple, the indirect AO is Issac Newton.

In reality, when measuring or testing the *effect* on B_i induced by A, we can not isolate the effect of the EXO, A_j, from that of the EO, A, on the direct AO, B_i. Instead, we have to take a holistic approach and consider the EO, A, the direct AO, B_i, and EXO, A_j, together as an entire system, under which we infer the true effect of the EO on the direct AO from the overall effect. We formally name this entire system, *a self-contained evaluation system (in short, SES)*. Upon the removal of the EO from the SES, we call the derived system the evaluation conditions (in short, EC). We illustrate the fundamental components of an SES in Figure 8.2.

If an EO has an indirect AO, a derived problem to be addressed is *how to infer the derived effect*. The same methodology mentioned above can be utilized to uncover the derived effect. We call the other objects that impact the derived effect, *the derived*

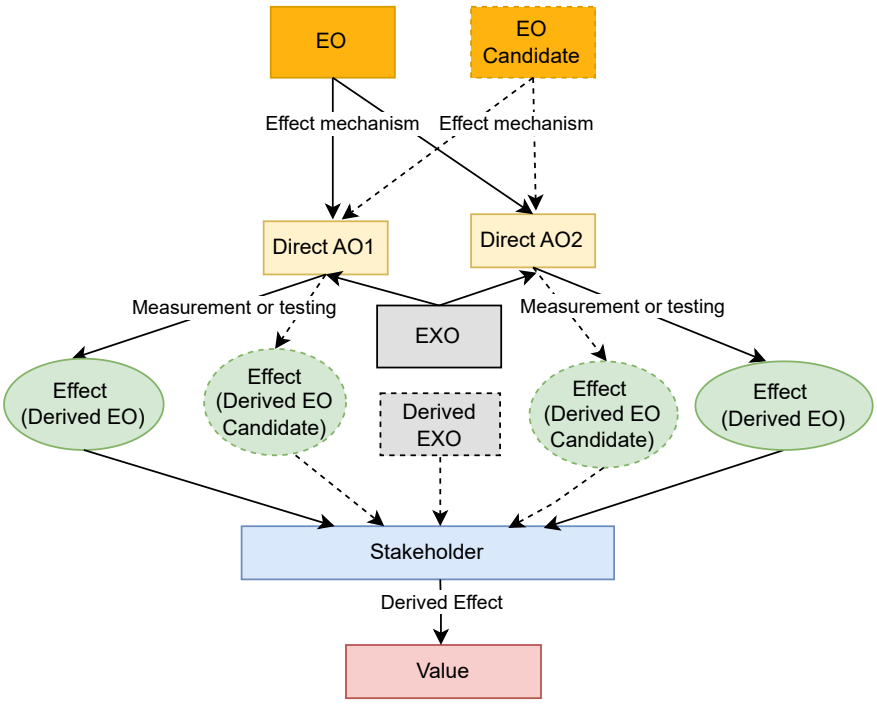


Figure 8.3: Essentially, the value is the derived effect on the stakeholder.

EXO. We isolate a derived SES, which includes the derived EOs, the indirect AO, and the derived EXO. Upon the removal of the derived DO from the derived SES, we call the derived system the derived EC. Within the context of this derived SES, we can infer the derived effect. However, it may include many direct or indirect AOs; the effect of the EO could be passed through several indirect AOs, and the chains could be long.

A *stakeholder* is an intelligent life or a system consisting of intelligent life that holds a stake of responsibility or interest in the EO. If an EO has a stakeholder, I call the effects of the derived EOs on the stakeholder, *the value of the EO*, as a stakeholder could make free choices among different EO candidates. The problem could be further stated as *how to infer the value of the EO*.

I formally call the effects of other EO candidates on the direct AOs *the derived EO candidates* when the stakeholder makes different choices. The same methodology that infers the derived effect still works, albeit with a slight difference, as other indirect AOs differ from the stakeholder in that they cannot make free choices. To infer the value of an EO on a stakeholder, a derived SES needs to include the derived EOs, the derived EO candidates, and the derived EXOs, under which the value could be inferred.

In summary, the essence of the evaluation is to *uncover an EO's effects and derived effects*.

A Typical Example on Drug Evaluation.

The First Step

When evaluating a drug, the first step is to identify a set of objects, e.g., a specific patient or a group of patients, on which differences can be measured or tested, depending on the presence or absence of that drug. The differences could be treatment outcome or cost.
In this case, the drug is the cause, a specific patient or a group of patients is the direct AO, and the measurable or testable difference in a specific patient or a group of patients is the effect induced by the drug. Often, identifying the mechanism of the effect of the drug is very challenging in biology.
For the treatment outcome, the EXO may include the working conditions, which may exert different pressures on the patient, and the living conditions, which may have different air or water quality.
In reality, we can not isolate the effect of the working conditions and living conditions. So, the SES contains the EO, the AO, and the EXO together, under which we reveal the effect of the drug.

The Second Step

In the drug evaluation case, the stakeholders could be patients' families, hospitals, or the government. The second step is to reveal the effect of the derived EO on the stakeholder. To address this issue, we need to isolate a derived SES, which includes the derived EOs, the derived EO candidates, the stakeholder, and the derived EXOs. Within the derived SES, we can infer the value of the EO.

8.3 Summary

This chapter presents the basic concepts of evaluation, including the cause, effect, and effect mechanism. I reveal that the essence of evaluation is how to uncover an EO's effects or derived effects.

Chapter 9

Evaluation Assumptions and Axioms

In this chapter, I present the assumptions and axioms of evaluation.

9.1 Assumption of Three Worlds

In this section, I elaborate on the assumption of the three worlds. The assumption of the three worlds states that there are three worlds: *the microscopic object world*, *the normal object world*, and *the free will world*, which are governed by consistent but different laws. Essentially, a law is a model.

A microscopic object world consists of *microscopic objects* at the scale of atoms and subatomic particles, which is governed by *the principles of quantum mechanics* [20, 122] [1].

Microscopic objects possess four properties, commonly referred to as *quantum effects*, which are fundamentally different from those of normal objects.

The first property is quantization, that is, energy, momentum, angular momentum, and other countable properties of microscopic objects are not continuous but come in discrete packets called "quanta."

The second property is wave-particle duality, that is, microscopic objects, like electrons and photons, exhibit both particle-like properties (e.g., having a location, colliding) and wave-like properties (e.g., interference, diffraction).

The third property is quantum superposition, that is, a microscopic object, or so-called quantum system, like an electron or photon, can exist in a combination of multiple possible states at the same time until it is measured. Upon measurement, it "collapses" into just one of those possible states.

Schrödinger's cat [169] is a famous example of microscopic objects, as being simultaneously both alive and dead until the box is opened and observed.

The fourth property is quantum entanglement, that is, two or more microscopic objects, like particles, can become linked or entangled in such a way that the quantum

[1] Mr. Chenxi Wang also contributed to the investigation of the principles of quantum mechanics.

state of one particle is instantly correlated with the state of the other(s), no matter how far apart they are. Measuring one instantly determines the state of the other.

Some Examples on the "Strange" Properties of Microscopic Objects

Electrons Jump Between Energy Levels

Electrons in an atom can only occupy specific energy levels (orbits). They jump between these levels by absorbing or emitting a quantum of light (a photon) [61].

Double-Slit Experiment

In the Double-Slit Experiment [30], a single electron fired at two slits behaves like a wave and creates an interference pattern on a screen, as if it passed through both slits simultaneously. Yet, it hits the screen at a single point like a particle.

Photons in an Entangled Pair

When the polarization direction of one photon in an entangled pair [23] is measured and determined, the polarization direction of the other photon immediately becomes correlated in a non-local and instantaneous manner [220].

Quantum Tunneling Effect

In semiconductor devices, electrons can traverse a potential barrier that is higher than their own energy, a phenomenon known as the quantum tunneling effect [216]. Even when an electron's energy is lower than the height of the potential barrier, there is still a certain probability that it will appear on the other side of the barrier.

A normal object is the object at the normal scales without free will. A normal object world consists of normal objects, which are determined and governed by *the principle of cause and effect*. This principle implies that an object as a cause will induce an effect on the other objects that can be objectively measured or tested. Please note that an automatic object, which we defined in Section 3.3, is a normal object, though it owns interrogation capacity and capability with different levels of complexity. Section 8.1 provides many examples of normal objects.

Quantum effects of microscopic objects reveal a universe far stranger and more probabilistic than that of the normal objects. At the smallest scales, certainty that dominated the normal objects gives way to probability distributions. Microscopic objects exhibit both particle-like and wave-like properties, exist in multiple states at once, and can be mysteriously connected across vast distances [20, 122]. These models of microscopic objects, while counterintuitive, are incredibly successful at explaining and predicting the

behavior of matter and energy at the atomic and subatomic level.

A free will world consists of the normal objects and the intelligent lives, and a free will world is governed by *both the principle of cause and effect and free will*. That is to say, an intelligent life not only receives the effects of causes, but also has the capacity and capability to make free and intentional choices

Typical examples of free will include love, writing, scientific research, art, voting, commodity purchasing, and college/program application in human society.

9.2 Evaluation Axioms

In this section, I present five evaluation axioms that are derived from the essence of evaluation, which I explain in Chapter 8. They focus on key aspects of evaluation outcomes, including true quantity (The first and second axioms), traceability of discrepancy (the third axiom), comparability (the fourth axiom), and estimate (the fifth axiom).

9.2.1 First Axiom of Evaluation Outcome

This axiom declares that *the essence of the evaluation outcome either uncovers an EO's effect or derived effect that carries inherent physical significance or is solely dictated by the value function.*

In nature, an evaluation metric refers to a base quantity, which indeed possesses physical significance, or other quantities that possess physical significance, or a combined quantity that is constructed using base quantities and other quantities.

There are two categories of evaluation metrics. In the first category, an evaluation metric carries inherent physical significance, revealing the effect of the EO on the direct AO, that is, the change of countable property of the direct AO induced by the EO.

In the second category, an evaluation metric reveals the effect of the derived EOs on the indirect AO or stakeholders. An EO can induce different effects on the same EO; that is, the EO's different quantities are affected. An EO can induce the same or different effects on different AOs. Those effects can be passed to the same indirect AO or stakeholder.

In this context, those effects are the changes in different countable properties of the AOs. The value function serves as a mechanism that maps base quantities and other quantities carrying physical meaning into a composite evaluation metric. The composite

evaluation metric does not necessarily carry inherent physical significance. Instead, it is defined as a value function that could be interpreted in a derived SES.

9.2.2 Second Axiom of True Evaluation

This axiom declares that *for an EO, when its SES is known, the effect of the EO on the direct AO possesses true outcomes; If an EO has an indirect AO and its derived SES is known, the derived effect on the indirect AO possesses true outcomes.*

In Section 8.2, I have formally defined what an SES is. Hereby, I further explain what I mean by referring to "when its SES is known." This proposition has a two-fold meaning. First, an SES is known, indicating that an SES and its components are known by the subject other than itself. Second, the effects of the EO and EXO, that is, the crucial components of the SES, on the direct AO can be measured or tested by the subject other than itself.

When revealing the effect of the EO, if we can isolate the SES, all relevant objects that have effects on the AO are identified, and their effects on the direct AO can be measured or tested. In this way, we can derive the true outcome of the effect.

Suppose the EO has an indirect AO. When uncovering the effect of the derived EO on the indirect AO, the same methodology can work.

Otherwise, the effect or the derived effect is undefined, and the subject other than the EO can not infer it.

9.2.3 Third Axiom of Evaluation Traceability

This axiom declares that *for the same (derived) EO, the divergence in the (derived) Effect can be attributed to disparities in (derived) ECs, thereby establishing evaluation traceability.*

This axiom focuses on the traceability of discrepancies in the effect. For the same EO, when the direct AO and the EXO are defined, the effect of the EO has a true outcome according to the First Axiom of Evaluation. The disparities in the overall effect of the EO can be rationalized as the consequence of variations in the ECs. In the absence of this axiom, the differences observed in the overall effect would be inexplicable, contradicting our scientific and engineering intuitions.

For the derived EO, the disparities in derived effects can be rationalized as the consequence of variations in the derived EC. The axiom holds for the same reason.

9.2.4 Fourth Axiom of Comparable Evaluation Outcomes

This axiom declares that *when different pairs of well-defined (derived) EO and direct (indirect) AO are equipped with equivalent (derived) EXO, their (derived) effects are comparable.*

It goes without saying that this axiom is related to the comparability of the evaluation outcomes. The equivalent EXO provides the same reference. When a pair of well-defined

EO and direct AO is equipped with the equivalent EXO, its evaluation outcomes possess true values.

Additionally, when two pairs of well-defined EO and direct AO are subjected to the equivalent EXO, their evaluation outcomes accurately reflect the effects of the EO on the direct AO under the same conditions, making them comparable.

When two pairs of well-defined derived EO and indirect AO are subjected to equivalent derived EXO, their derived evaluation outcomes accurately reflect the effects of the derived EO on the indirect AO under the same conditions, making them comparable.

9.2.5 Fifth Axiom of Consistent Evaluation Outcomes

This axiom asserts that *when uncovering the effect of a known (derived) EO on a known direct (indirect) AO using different samples from a population of (derived) EXOs, their (derived) effect consistently converges towards the true quantities.*

This axiom provides an estimate of the true outcomes of the (derived) effect under the population of (derived) EXOs. According to the Second Axiom of Evaluation, when a pair of the known EO and the direct AO is equipped with a population of known EXOs, the effect possesses the true outcomes. When a sample is taken from a population of EXOs, it serves as an approximation of the entire population. As a result, different samples yield consistent effect outcomes that converge towards the true outcomes of the entire population of EXOs. This convergence is influenced by the sample's ability, which is determined by the chosen sampling policy, to represent the underlying characteristics of the population accurately.

When a pair of a known derived EO and a known indirect AO is equipped with a population of the known derived EXOs, the derived effect possesses the true outcomes. The axiom holds for the same reason.

9.3 Summary

This chapter presented the assumption of three worlds and the five axioms of evaluation.

Chapter 10

Revisiting Interrogations

In this chapter, I revisit and redefine several concepts and clarify the unique positions and interrelationships of four interrogations: measurement, testing, reasoning, and evaluation.

10.1 Revisiting Several Concepts

As the essence of the evaluation is to uncover the effects of objects, this section will extend several definitions in Chapter 5.

A *proposition* is a testable statement about an object or *several objects with mutual influences*.

A *model* is a streamlined representation of an object or *several objects with mutual influences* [221, 224]. A model can manifest as a physical, mathematical, or other construct, e.g., a causal model. A valid model that can be verified objectively by a subject other than itself is a kind of *fact or truth*. A model about an object or *several objects with mutual influences* could be a hypothesis under test, which is defined in Section 3.6.

A *fact or truth* is a proposition or a model about an object and *several objects with mutual influences* that can be proven true or verified objectively by a subject other than itself.

For instance, the fact that water boils at 100 °C under standard atmospheric pressure describes an object (water) and its interaction with another object (air) and can be objectively verified by the observer (the subject) using a thermometer and a barometer.

A *test oracle* is the ground truth or facts about an object or *several objects with mutual influences*. Additionally, for artifacts, a test oracle could be the intended behavior expected for an object or several objects with influences.

For instance, in a system with multiple interacting objects, such as a temperature sensor (Object A) and a cooling fan (Object B), the test oracle defines the expected behavior between them: when the temperature exceeds 50 °C, the fan must switch to high speed ($\geq$3000 RPM) within one second, and when the temperature falls below 30

°C, it must return to low speed (≤1500 RPM).

A *test case* is a predefined interrogation condition for testing that is designed and implemented for a test oracle and is ready for executing an object or *several objects with mutual influences* under test to verify whether the actual outcomes are consistent with the mandated or expected results defined by the test oracle.

For instance, a corresponding test case may be designed for the test oracle by raising the temperature from 25°C to 45°C and expecting the fan to reach at least 3000 RPM within one second.

10.2 Redefinition of Testing and Reasoning

Based on the extended concepts, I redefine testing and reasoning.

Testing is a verification process of running test cases to determine whether a proposition or a model of an object or *several objects with mutual effects* conforms to the test oracle through comparing their outcomes [224, 223].

The proposition or a model states what is expected for an object or several objects with mutual influences under an interrogation condition. The test oracle mandates the reference outcome under the reference interrogation conditions. The essence of testing is to compare the interrogation outcomes to those of the test oracle.

Being testable or testability means a proposition or a model of an object or several objects with mutual influences can be falsifiable through comparing its outcome with that of test oracles.

Reasoning is a capacity, capability, and mental process that generates new propositions about an object or *several objects with mutual influences* based on the propositions, models, and other interrogation outcomes.

The definition of measurement remains unchanged.

10.3 The Unique Position of Four Interrogations

In the epistemic hierarchy of Evaluatology, measurement, testing, reasoning, and evaluation represent four fundamental interrogations through which intelligent lives explore the unexplored world and their unknown lives, and build massive knowledge systems.

Measurement answers “how much”, attributing values to countable quantities of objects; testing answers “whether”, determining conformity to the test oracle through verification and falsification; evaluation answers “why” in terms of how an object influences another one, or mutual influence between objects; reasoning answers “why” in terms of the the underlying logical mechanisms that connect causes to their effects. Together, these four interrogations form a complete cognitive cycle—from observation, to

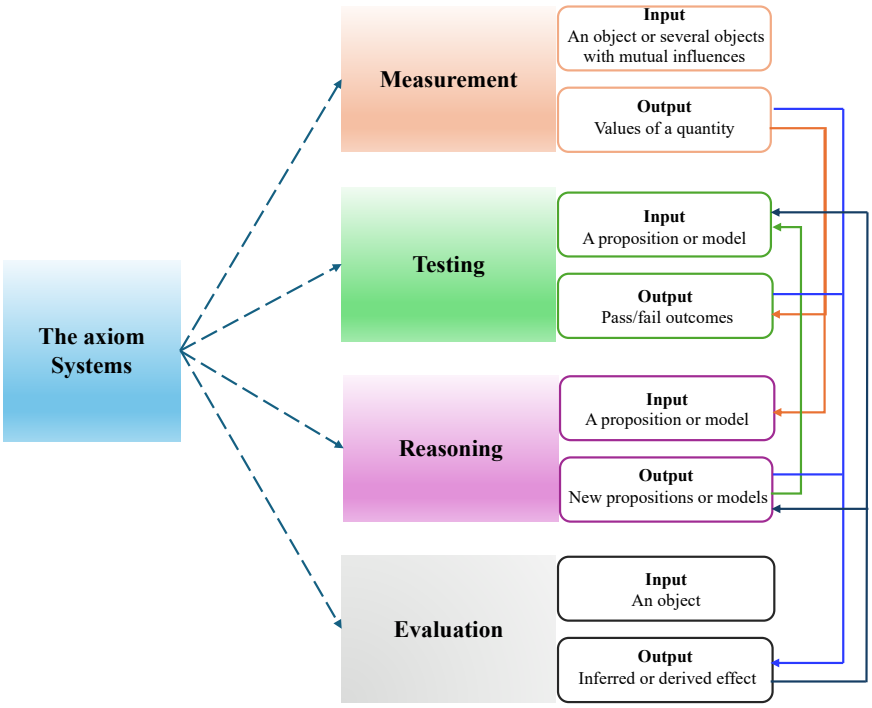


Figure 10.1: Interrelations of the four interrogations.

validation, to explanation [1].

10.4 Interrelations of the Four Interrogations

The interrelations of the four interrogations are depicted in Figure 10.1 [2]. Measurement, testing, reasoning, and evaluation are built on an axiom system that is accepted as the truth without proof.

Measurement attributes the value to the quantity of an object or several objects with mutual influences, based on which the subject generates propositions or models that serve as the input for reasoning, or compares the outcomes of testing cases to those of test oracles to draw a testing conclusion.

Through reasoning, the subject could generate new propositions or models about an object or several objects with mutual influences, which can only be validated by testing.

Evaluation relies upon measurement, testing, and reasoning to infer the true effect of an object. Evaluation reveals the effect of objects, which could generate the propositions or models that serve as the input for testing or reasoning.

[1] Dr. Fanda Fan's work provided a basis for this passage.
[2] Dr. Wanling Gao contributed to this figure based on the discussion with Dr. Lei Wang and me.

Testing could generate new propositions about an object or several objects with mutual influences. From this perspective, testing not only validates reasoning but also provides new input for reasoning.

10.5 Summary

This chapter expands the definitions of testing and reasoning to encompass several objects with mutual influences. Also, I clarify the unique positions and four interrelationships among the four interrogations: measurement, testing, reasoning, and evaluation.

Chapter 11

Universal Evaluation Concepts

In this chapter, I delve deeper into various evaluation cases across diverse disciplines, with the primary aim of enhancing our comprehension of the evaluation. For the sake of convenience, henceforth, within this section, each case study shall be assigned a unique case ID for differentiation purposes.

11.1 A Computer Science Example: The CPU evaluation

In the first case, which we will refer to as Case One, an organization is in the process of acquiring a critical component of the computer, the CPU. To make an informed decision, the organization evaluates various CPU options by executing its applications on a computer built upon the corresponding CPU. During this evaluation, the organization will collect extensive data on performance and energy efficiency.

In any evaluation process, we refer to the object, individual, or system under evaluation as an EO. As discussed in 8.1, when we uncover the effect of the EO, the EO is the cause object or cause. In the context of Evaluatology, we do not distinguish between EO or cause unless explicitly stated.

A direct AO is defined as *a minimal system on which the effect of the EO can be directly measured or tested* in Section 8.2. Whether and how to measure or test the EO's effect will determine the scope of the direct AO. In Case One, the prospect of measuring the mere properties of an object, such as its weight and power consumption, possesses a certain degree of utility. Nevertheless, such measurements fall considerably short of meeting stakeholders' evaluation requirements.

If the effect of the CPU is in terms of the running time of a typical workload, it can only be directly measured or tested on a minimal system built on that CPU, including the OS, memory, and disk, as shown in Figure 11.1, which is the direct AO. Significantly different from the remaining cases we discussed later, the EO is a component of the direct AO in this case.

For an EO and a specific AO instance, the EXO also impacts the overall effect.

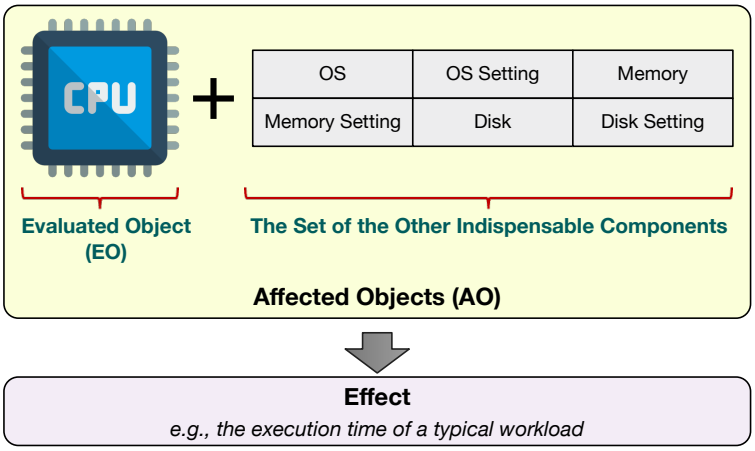


Figure 11.1: For a CPU, its AO is a minimal system built on it, of which the CPU is a component of the AO.

Figure 11.2 showcases how the overall effect is measured on a direct AO instance built on a specific CPU, with the impact of the varying EXO.

After obtaining different dimensions of the effect of the EO, a further question remains: *How does a stakeholder determine different CPU options when they exhibit varying levels of performance across different applications?*

In Case One, the stakeholders could be an organization seeking to acquire the computers, a specific user with different requirements, the designers responsible for creating the CPU specifications, and the producers who manufacture the CPUs.

Based on the collected data, the organization will then formulate an explicit or implicit function that can map different dimensions of the effects of EO on the direct AO to one or several composite evaluation metrics. In the rest of this article, I refer to this function as *a value function*. I interpret the EO's merit or worth [172, 174, 175, 176] as the value this function maps.

According to the discussion in Section 8.2, *the value of an EO reflects the effects of derived EO on the stakeholder*. Please note that the effects of the EO can be propagated to multiple direct and indirect AOs, ultimately reaching the stakeholder.

Figure 11.3 [1] illustrates a typical CPU stakeholder. Their default direct AO component configuration excludes the CPU. The stakeholder operates a business where application execution speed directly impacts profitability—faster performance yields higher revenue. However, increased CPU speed comes at the cost of higher power consumption.

When a user compares two CPUs, A and B, their performance profiles exhibit distinct trade-offs: A delivers higher computational speed at the expense of increased power consumption, while B prioritizes energy efficiency with reduced speed. Since A outperforms B in speed, it generates greater revenue for stakeholders. Consequently, the economic impact of the derived EO (e.g., revenue vs. cost trade-offs) can be quantified

[1] Mr. Chenxi Wang contributed to this figure based on the discussion with me.

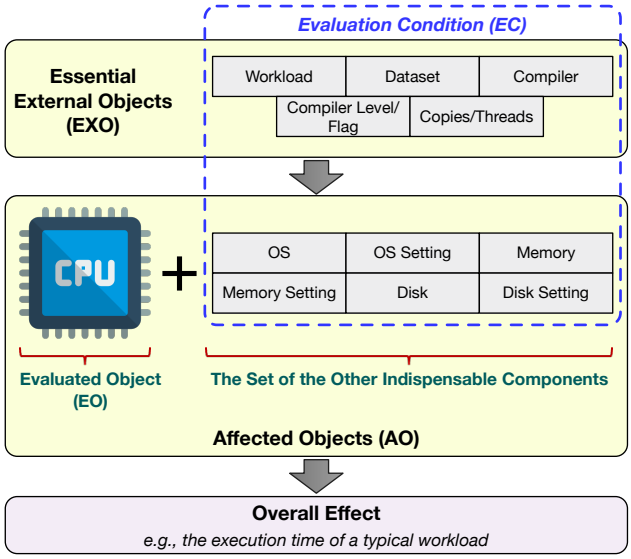


Figure 11.2: A typical CPU example showcases how the evaluation outcome (overall effect) is measured on an AO built on a specific EO (CPU), with the impact of the varying EXO.

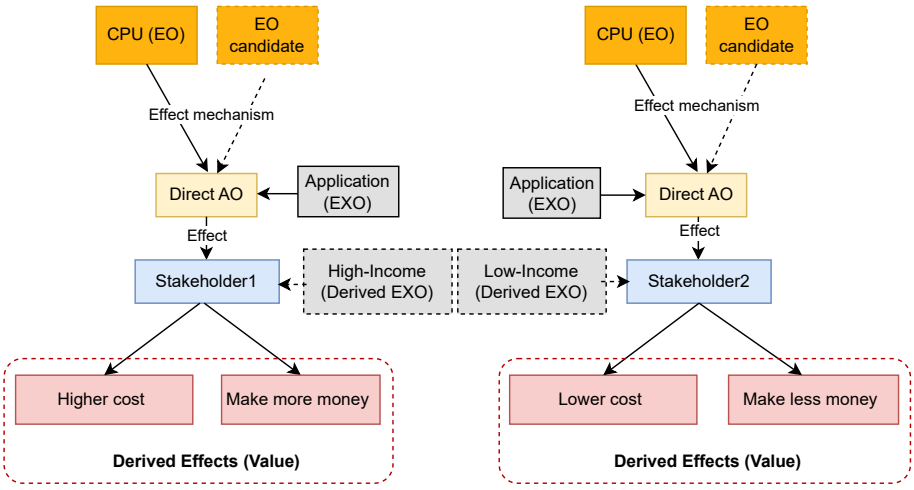


Figure 11.3: For a typical stakeholder, how the value is assigned to the CPU.

for stakeholders.

The stakeholder's valuation of these CPU candidates further depends on the stakeholder's derived EXO, such as the financial context and revenue expectations, which influence how stakeholders assign the value to different CPU options.

11.2 A Physics Example: Issac Newton's Apple Tree

It is a well-known legend that an apple tree at Cambridge University—later immortalized as the "Newton Tree"—inspired Isaac Newton, a once-in-a-millennium genius, to ponder why an apple falls from the tree (Case Two). I argue that Case Two exemplifies the inverse problem of evaluation, which I formally termed "de-evaluation" in Section 12.1.3.

Newton observed an apple falling from the tree and pondered the underlying cause. In Case Two, the direct AO is unambiguously the apple itself. Isaac Newton could measure numerous quantities of the direct AO, which collectively reflect the overall effects induced by myriad objects, including the unknown EOs. However, Newton faced three formidable challenges in addressing this phenomenon.

- What primary EO candidates induced the apple to fall from the tree?
- What is the effect mechanism through which the primary EOs act upon the direct AO?
- How to infer the effect of the primary EO?

Issac Newton could guess and identify several EO candidates and design different SESes, through which he could uncover the primary EO and its effect mechanism that induces the apple to fall from the tree.

Case Two has three unique characteristics. First, it is very challenging to isolate an SES. Even Newton was bold and daring; he could include the Earth, the Sun, the Moon, the wind, and the air into the SES. There could be other invisible or mysterious objects, as shown in Figure 11.4.

Second, it is almost impossible to change different settings in this SES. However, Issac Newton is a genius; he could design many extreme SESes to isolate other objects, which was later perfectly achieved by Henry Cavendish.

For instance, an orange, a pear, or a perch could substitute for the apple, demonstrating that the effect mechanism is independent of the AO. An EC populated with a vacuum would eliminate the influence of wind and air, while an EC populated with a vacuum at different physical locations on Earth would reveal that the effect mechanism varies with geographical location.

Henry Cavendish's experiment [31, 108] to measure the gravitational constant (G) in 1797 provided a very simple SES through many intricate experimental skills.

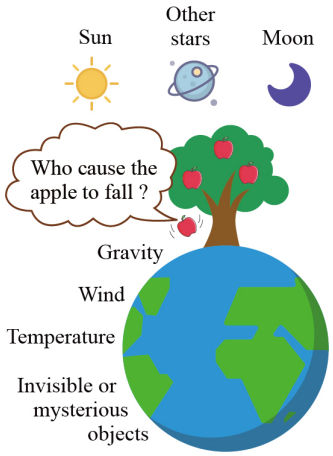


Figure 11.4: For identifying the cause that induces an apple from a tree, it is very challenging to isolate an SES.

How Henry Cavendish Designed a Very Simple SES to Measure the gravitational constant (G)

Apparatus.

- A torsion balance consisting of a six-foot wooden rod suspended from a wire, with two small lead spheres (1.6 pounds each) attached to its ends.
- Two larger lead spheres (350 pounds each) were placed near the smaller ones, separated by about nine inches.
- A mirror attached to the wire reflected a light beam onto a scale, amplifying tiny torsional movements.

Method.

- The gravitational attraction between the large and small spheres caused the rod to twist slightly, deflecting the light beam.
- By measuring this deflection, Cavendish calculated the force between the masses and derived G using Newton's inverse-square law.

Mitigation of External Objects' Effects to Isolate the SEC.

Shielding Earth's Gravity

- Symmetrical setup: The experiment balanced Earth's gravitational pull on both sides of the torsion balance, canceling its net effect.
- Local mass symmetry: Large lead spheres were placed symmetrically around the small spheres to minimize asymmetric gravitational gradients.

Isolation from nearby objects

- Sealed enclosure: The apparatus was housed in a closed room to eliminate air drafts and reduce temperature variations
- Non-magnetic materials: Lead was chosen to avoid magnetic interference from external objects like magnetic fields

key findings

- Cavendish determined the gravitational constant as G $6.754\times10^{-11}\,\mathrm{N\cdot m^2/kg^2}$, remarkably close to modern values (6.6732×10^{-11}).
- From G, he calculated Earth's average density (5.481 times water's density) and mass.

Sources of Error in Cavendish's Experiment

Environmental Disturbances

- Air currents: Even slight air movements could cause false torsion balance deflection.
- Temperature fluctuations: Thermal expansion of the suspension wire or lead spheres might alter the measured torque.

Instrumental Limitations

- Non-ideal masses: The lead spheres were not perfect point masses, introducing deviations from Newton's inverse-square law.
- Torsion wire imperfections: Variations in wire elasticity or manufacturing defects could affect the torsion constant.

Measurement Uncertainties

- Optical alignment: Mirror misalignment or scale parallax errors in tracking light spot displacement.
- Periodic oscillations: Damping incomplete torsional oscillations led to timing errors in period measurements.

Through the design of simple SESes, Issac Newton could speed up the process that drew close to the Law of Universal Gravitation; however, it is still very challenging. Through the work of Issac Newton, we know the EO is the Earth, and the effect mechanism is gravitation. However, if we aim to fully understand the effect of the Earth on the other objects, there are many open issues.

Third, uncovering the effect of Earth's gravity on the apple is a purely Physics problem; discussing its value is inherently meaningless. However, if Earth's influence extends to the apple's sweetness—a property that plant owners or end-users do care about—then it becomes meaningful to evaluate the Earth's value in this context. This is particularly relevant if we consider planting apple trees on the moon or other EO candidates, where the Earth's role as an EO could be changed.

Fourth, even in uncovering the effect of the Earth on the apple, there are many different perspectives. For example, the Earth also provides indispensable living environments for planting an apple tree.

11.3 An Artificial Intelligence Example: Evaluating an AI algorithm

In Case Three, the objective is to evaluate an AI algorithm, specifically focusing on an Image Classification task as a case study. Real-world images are collected and annotated with accurate labels, such as a cat or a dog. A portion of these images is randomly selected to construct the training, validation, and testing datasets based on a designated percentage. To assess the image classification algorithm, that is, the EO in this case, it must be implemented on a computer system utilizing a specific programming framework, such as PyTorch or TensorFlow.

During the evaluation process, the testing data is provided to the algorithm, which generates an output. This output is then compared to the ground truth: the labels.

Additionally, measurements are collected for each run of the evaluation process.

The fundamental evaluation process in Case Three bears a resemblance to that of Case One. However, there exist three notable distinctions. First, the EO, i.e., the algorithms, must be implemented on a computer. This case has two-fold implications. On the one hand, the EO has different implementations, implying that different EO instantiations have a diversifying effect. On the other hand, within a complex system featuring diverse implementations, the EO instantiation is initiated, and these varied implementations significantly influence the effects of the EO instantiation.

Secondly, the EC in Case Three differs. The presence of a dataset labeled with ground truth constitutes a vital aspect of an EC. A dataset labeled with the ground truth represents a specific instance of a problem or task. Regrettably, in Case Three, it is challenging to define a problem or task mathematically. Hence, diverse problem or task instances are devised in an ad-hoc manner, for example, selecting images randomly to form training, validation, and test datasets based on a predetermined percentage. Given that image selection relies on professional expertise, varying professional expertise will inevitably introduce distinctions among ECs.

Thirdly, in Case Three, upon feeding the algorithm with the testing data, it generates an output that is then compared with the ground truth, also known as the test oracle. Apart from measurements, there exist other forms of activities in the evaluation process, namely testing, as expounded upon in Chapter 5.

Fourth, the direct AO and EXO encompass significantly more diverse and complex components. For example, if stakeholders prioritize algorithmic accuracy, the direct AO includes the dataset labeled with ground truth, the algorithm implementation, the operating system, the processor, and the memory, while the EXO includes the compiler, the programming framework, like PyTorch.

11.4 An Chemistry Example: Lavoisier's Combustion Experiment

Case Four [2] is a famous chemistry example: Lavoisier's combustion experiment. Lavoisier's combustion experiment is considered a milestone in the history of chemistry, as it revealed the crucial role of oxygen in the combustion process [109].

Lavoisier's classic Combustion experiment, central to his investigation of the transformation of matter and the composition of air, was both ingeniously designed and meticulously executed.

In this experiment, as shown in Figure 11.5, mercury was placed in a distillation flask and heated, with a mercury trap below to capture the evaporated mercury vapor. The vapor was then guided through a pipe into a separate container, where it cooled and was re-liquefied via a condenser. The experiment was conducted in two phases:

[2] Dr. Lei Wang authored this subsection.

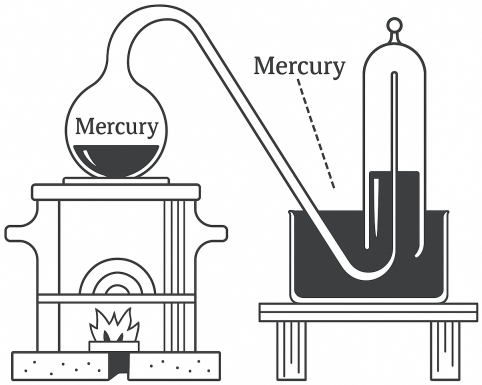


Figure 11.5: Overview of Lavoisier's classic combustion experiment.

In the first phase, heating continued for twelve days, during which mercury reacted with components in the air to form red mercuric oxide. The volume of gas in the system contracted by approximately one-fifth, and the total mass of mercury and mercuric oxide remained unchanged, providing preliminary support for the principle of mass conservation. In the second phase, mercuric oxide was subjected to intense heat in the distillation flask, decomposing into a gas that could support combustion and respiration. Notably, the volume of this gas precisely filled one-fifth of the space previously reduced, while the remaining four-fifths of the gas exhibited no such properties.

Through this experiment, Lavoisier not only demonstrated the reversibility of mercury vaporization and oxidation —confirming that mass was not lost but merely transformed —but also conclusively proved that air is composed of "combustible oxygen" (approximately one-fifth) and "inert nitrogen" (approximately four-fifths).

This groundbreaking discovery effectively tested and invalidated the phlogiston theory, establishing the foundational principles of mass conservation and oxidation reactions and laying the cornerstone for modern chemistry.

The goal of this experiment is to identify the primary cause object that induces combustion. If we treat Lavoisier's Combustion Experiment as an evaluation, the EO is some mysterious substance to be determined.

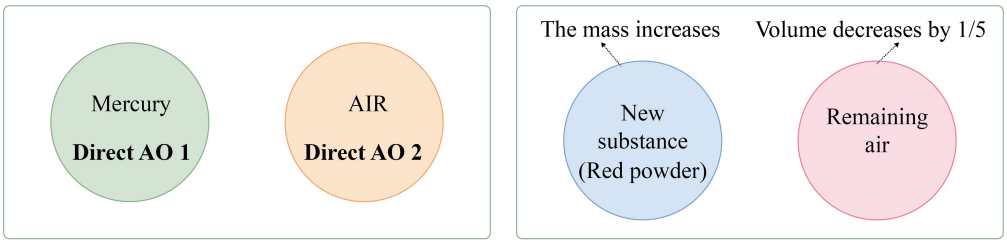
The direct AOs are any substances that can be burned or calcined, such as mercury, phosphorus, and sulfur, and we can directly measure quantities of direct AOs, such as mass, gas volume, and temperature. Before Lavoisier's combustion experiment, people believed that the mysterious substance was "phlogiston" and the EO (phlogiston) was a part of these direct AOs.

Antoine Laurent Lavoisier designed various SESes centered around the closed space, which isolated the effects of other objects on direct AOs.

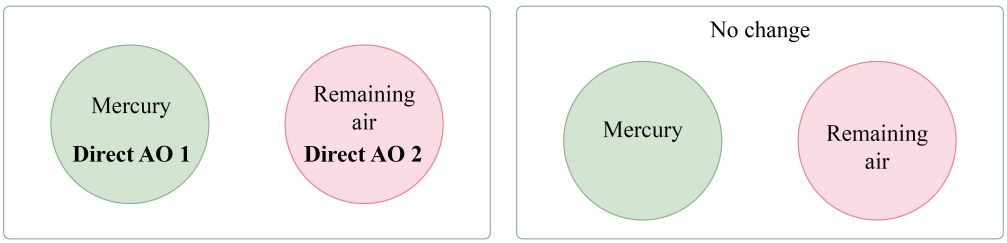
He sealed mercury (the direct AO) in a retort and heated it, observing changes in the gas volume within the closed space and collecting the gases produced by combustion for analysis.

The key steps included:

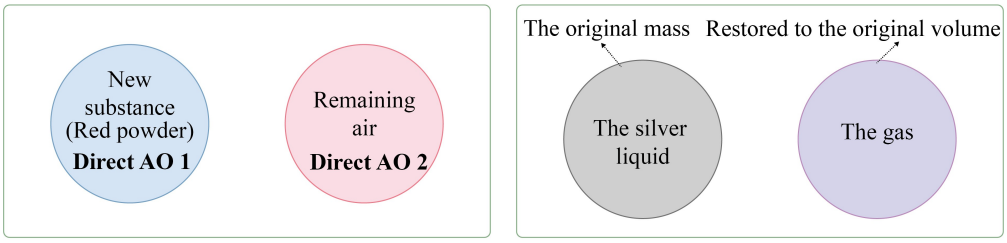
- *Heating mercury*: Sealed in a glass vessel, mercury formed red calx (HgO) while the air volume decreased by roughly one-fifth.
- *Gas analysis*: The remaining gas (later called *azote*, i.e., nitrogen) failed to support life or combustion.
- *Decomposition of calx*: When heated, HgO released a highly respirable gas (oxygen), and the gas volume returned accordingly.



(a) Phase one



(b) Phase two



(c) Phase three

Figure 11.6: Antoine Laurent Lavoisier's experiments: The changes after heating in three phases.

Significantly different from the other cases, Case Four has a unique characteristic: a

closed space could perfectly isolate the other objects' effect; however, though in the same closed space, the direct AOs and EXO change in different phases as shown in Figure 11.6.

In the first phase, there are two direct AOs: the mercury (silver liquid) and the air, consisting of unknown objects at his time. After heating, the mercury changes into a new substance (red powder), with the mass increased; the air volume decreased by about one-fifth.

In the second phase, there are two direct AOs: the mercury (silver liquid) and the remaining air from the first step. After heating, there is no change, and no combustion is observed.

In the third phase, there are two direct AOs: the new substance (red powder) and the remaining air produced in Step One. After heating, the red powder changed into a silver liquid with the original mass and released a highly respirable gas. Finally, the gas was restored to the original volume.

By comparing the SESes in different steps while using the closed space to isolate the other objects' effect (the control as we discussed in Part IV), Antoine Laurent Lavoisier induced that combustion depended on a reactive component in air (later named *oxygen*) and that, by volume, it accounted for about one-fifth of air. He ultimately overturned the prevailing phlogiston theory, which claimed combustion was due to the release of "phlogiston" and established the scientific theory that "combustion is the oxidation of substances by oxygen."

Additionally, he used other combustible materials (direct AO), such as phosphorus and sulfur, to verify the universality of the combustion mechanism.

He also revealed the effect of the mechanism in combustion as follows.

$$\text{Combustion: } 2\text{Hg} + \text{O}_2 \xrightarrow{\Delta} 2\text{HgO} \quad \text{(silver liquid} \rightarrow \text{red powder)}$$

$$\text{Decomposition: } 2\text{HgO} \xrightarrow{\Delta} 2\text{Hg} + \text{O}_2 \uparrow \quad \text{(red powder} \rightarrow \text{silver liquid + gas)}$$

Lavoisier showed:

- Weight gain in combustion came from oxygen absorption, not "phlogiston" release.
- Air comprised about 20% oxygen (vital for combustion) and about 80% nitrogen.

Despite contemporary debate over priority and interpretation, Lavoisier's use of precise balances and careful mass accounting cemented the oxygen theory of combustion. His experiments demonstrated that combustion is a chemical reaction between substances and oxygen, fundamentally changing our understanding of chemical processes.

His work introduced:

- *Law of Conservation of Mass*: The total mass in a reaction remains constant.
- *Oxidation Theory*: Combustion is the oxidation of substances by oxygen.

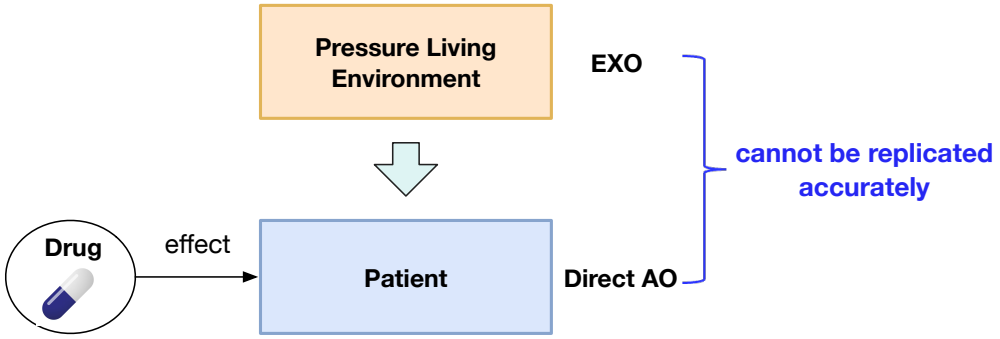


Figure 11.7: In the case of the drug evaluation, we can not replicate the direct AO and EXO very accurately.

11.5 A Pharmacology Example: Drug evaluations

The Fifth Case (Case Five) involves evaluating a drug. In this case, the EO refers to a specific drug, while the direct AO encompasses the participants (human beings) targeted by these interventions. Figure 11.7 shows the composition of the SES.

When comparing Cases Five with Cases One and Three, we have made three observations. First, the direct AO is independent of the EO. Second, the overall effect can be measured or tested on the direct AO, with the contributions of EXOs, e.g, the pressure exerted on the patients, and the environment in which the patients live.

The last but not least observation revolves around the distinctions of the EXO and direct AO found in Cases One, Three, and Five. In Cases One or Three, the direct AO and EXO can be well-defined; that is to say, we can replicate an EXO or direct AO very accurately. However, in Case Five, we can not replicate an EXO or direct AO very accurately. For example, we can not replicate a patient accurately, not to mention other critical factors included in the EXO. Instead, there is substantial variability in EXO and direct AO in Case Three.

11.6 A Social Science Example: Policy evaluation

The Sixth Case (Case Six) involves evaluating a policy aimed at addressing drug addiction within a community. In this case, the EO refers to a policy aimed at addressing drug addiction intervention. When comparing Cases Six with Cases Three, we have made three observations.

Firstly, they have a resemblance. The policy may have different instantiations, like that of the algorithm. The EO instantiation has a significant impact on its effect, as same to that in Case Three. However, in Case Six, the EO instantiations can not be controlled accurately, differing significantly from those of the algorithm. For the latter, the instantiations of an algorithm can be replicated completely and accurately.

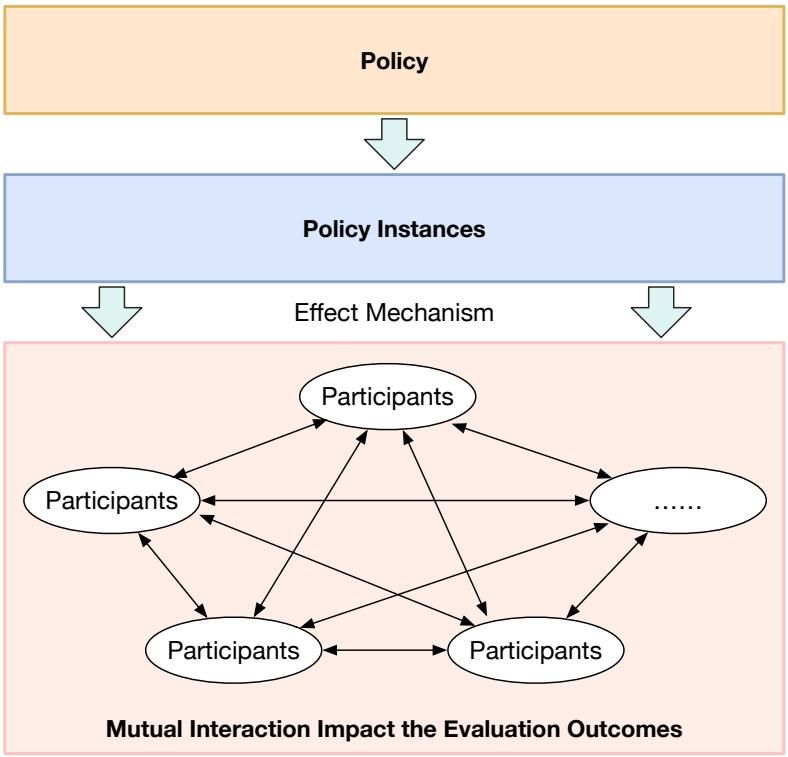


Figure 11.8: In the policy evaluation, the interaction of direct AO significantly impacts the evaluation outcomes.

Second, the EXO, which impacts the overall effect, is more difficult to identify in Case Six than in Cases Five or Three. For example, the participants' attitude towards the policy may dominate the evaluation outcomes.

Third, the interactions between different participants, the direct AOs, also impact the evaluation outcome significantly, as shown in Figure 11.8. In Cases One, Three, and Five, each direct AO is independent from the other; however, in Case Six, each direct AO is dependent on the others.

11.7 An Biology Example: Gregor Johann Mendel's Peas

Biological traits, such as human height and body shape, a cat's straight or curly fur, and the sweetness or waxiness of corn, represent the observable variations among biological individuals. What determines these biological traits? Why do the offspring of a tall-stemmed pea plant and a short-stemmed one all exhibit tall stems? And why do some offspring of tall-stemmed pea plants display short stems? The SES of the inheritance of natural traits is shown in Figure 11.9.

The answers to these questions were provided by Gregor Johann Mendel. Through

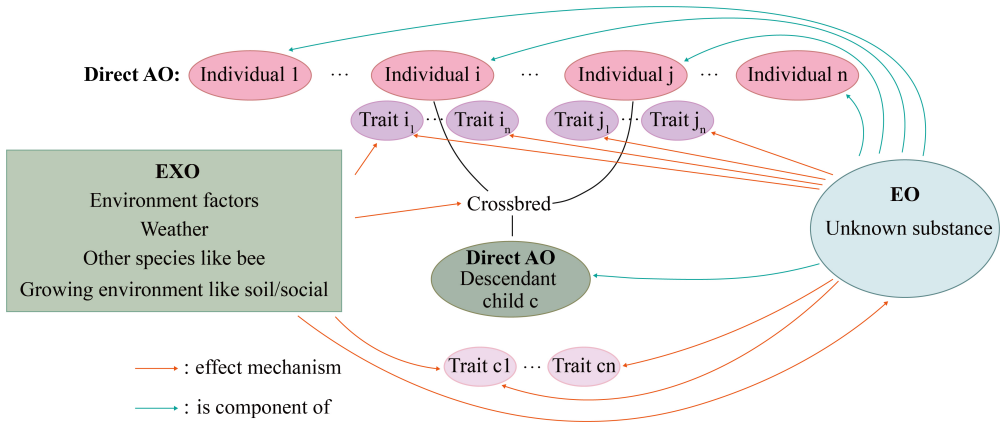


Figure 11.9: SES of the inheritance of natural traits, there is a complex relationship among the EO, the direct AO, and the EXO.

meticulously designed pea experiments (Case Seven [3]), Gregor Johann Mendel formulated the classic laws of segregation and independent assortment in genetics [120].

Different from other cases, Case Seven has a unique characteristic. When two peas are crossbred, both their gene (two EOs) are the cause, while the direct AO are those peas and their descendant. Figure 11.10 shows the complex relationship among the EC, the direct AO, and the EXO.

Mendel's choice of the garden pea (direct AO) was strategic and critical to his success:

- *True-Breeding Strains:* Peas can self-fertilize, and Mendel identified varieties that, when self-fertilized, consistently produced offspring identical to the parent for specific *traits* (e.g., seed shape). These were his *P generation* (parental generation).
- *Distinct, Heritable Traits:* He selected seven traits that exhibited clear-cut, contrasting *traits* (see Table 11.1), such as Round vs. Wrinkled seeds. There were no intermediates.
- *Controlled Crosses:* The physical structure of the pea flower allowed Mendel to easily prevent self-pollination and perform *cross-pollination* (*hybridization*) between selected parents.
- *Short Generation Time:* Multiple generations could be studied within a reasonable time frame.

Mendel initially spent several years allowing peas to undergo self-pollination, thereby selecting strains that could stably inherit specific traits (with consistent offspring phenotypes), known as "homozygous". For instance, the offspring of tall-stemmed pea plants would always be tall-stemmed. His initial experiments focused on a single trait at a time, a *monohybrid cross*.

[3] Mr Chenxi Wang authored this section.

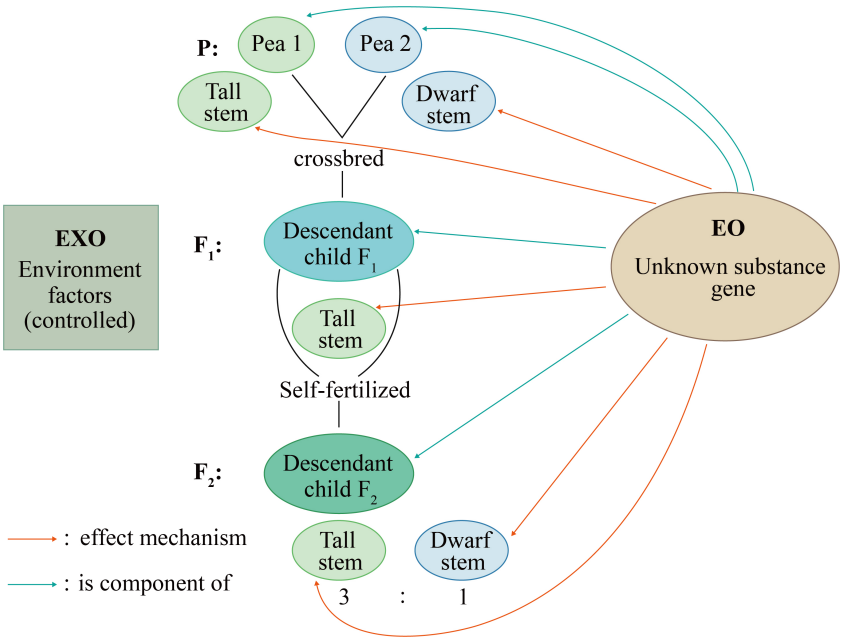


Figure 11.10: SES of Gregor Johann Mendel's Peas experiments [120], there is a complex relationship among the EO, the direct AO, and the EXO.

Trait	Dominant Trait	Recessive Trait
Seed Shape	Round	Wrinkled
Seed Color	Yellow	Green
Flower Color	Purple	White
Pod Shape	Inflated	Constricted
Pod Color	Green	Yellow
Flower Position	Axial	Terminal
Stem Length	Tall	Dwarf

Table 11.1: The seven pea plant traits studied by Mendel

- He crossbred two true-breeding (homozygous) pea plants that differed in one trait (e.g., Tall vs. Dwarf).
- He collected the seeds and grew the plants—this was the *first filial generation (F_1)*.

Observation One: The F_1 offspring always resembled only one of the parents. For example, the cross between a pure Tall and a pure Dwarf plant produced *all* Tall offspring. The trait that appeared in the F_1 generation he called the *dominant* trait (Tall). The trait that seemed to vanish was the *recessive* trait (Dwarf).

Mendel then allowed the F_1 plants to self-fertilize and produce the *second filial generation (F_2)*.

Observation Two: The recessive trait, absent in the F_1 generation, reappeared in the F_2 generation! Furthermore, Mendel counted the numbers of each type and found a consistent ratio of approximately *3 Dominant: 1 Recessive*.

To explain these results, Mendel proposed a revolutionary model:

- *Particulate Factors:* Inheritance is controlled by discrete "factors" (now called *genes*) that are passed unchanged from parents to offspring.
- *Alleles:* For each trait, an organism inherits two factors, one from each parent. These different forms of a gene are called *alleles* (e.g., a Tall allele and a Dwarf allele).
- *Dominance/Recessiveness:* In a heterozygote (with two different alleles), the *dominant* allele determines the organism's appearance, while the *recessive* allele has no noticeable effect.
- *The Law of Segregation:* The two alleles for a heritable trait segregate (separate) from each other during the formation of gametes (sex cells), so that each gamete carries only one allele for each trait.

Mendel then asked: What happens when two traits are inherited simultaneously? He performed a *dihybrid cross* between parents that differed in two traits—for example, a plant true-breeding for Round Yellow seeds ($RRYY$) and one true-breeding for Wrinkled Green seeds ($rryy$).

- *F_1 Generation:* All offspring were $RrYy$ and had Round Yellow seeds.
- *F_2 Generation:* When the F_1 plants self-fertilized, the F_2 generation showed *four phenotypes* in a consistent ratio:

9 Round Yellow (1 RRYY + 2 RRYy + 4 RrYy + 2 RrYY) : 3 Round Green (1 RRyy + 2Rryy) : 3 Wrinkled Yellow (1 rrYY + 2 rrYy): 1 Wrinkled Green (1rryy)

The 9:3:3:1 ratio was a dihybrid ratio that could be derived from two independent monohybrid 3:1 ratios ((3:1) × (3:1) = 9:3:3:1). This led to his *Law of Independent Assortment*.

- *The Law of Independent Assortment:* Genes for different traits assort independently of one another during gamete formation. The allele a gamete receives for one gene does not influence the allele received for another gene.

In Mendel's pea experiments, the EO refers to genes, the direct AO denotes peas, which can be substituted with any biological individuals, the effect refers to the traits exhibited by the peas, and the effect mechanism pertains to the laws of segregation and independent assortment.

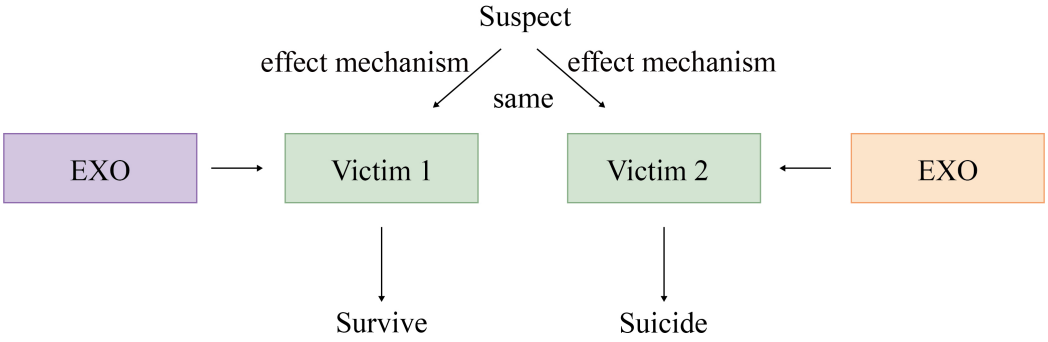


Figure 11.11: A suspect uses violent language with the same intensity to insistently intimidate two victims. The overall effects on the two victims are significantly varied.

11.8 Scientific and Technology Achievement Evaluations

Case Eight involves evaluating a specific scientific or technological achievement. The EO is obvious, including but not limited to a specific idea, an article, a patent, or a technology product.

Compared with other cases, Case Eight has two significant distinctions. First, in Case Eight, the direct AO is not limited to a single kind of object. For example, it could be the other scientists or engineers, other achievements, the industry, or even the society. Instead, in the other cases, the direct AOs are basically limited to a single kind of object. Second, it is almost impossible to perform an experiment to investigate the effect of the EO. We have to take a retrospective instead of a prospective approach.

11.9 A Legal Liability Example: Determining Accountability for Suspects.

In Case Nine, we presume there is a suspect who uses violent language with the same intensity to insistently intimidate two victims, and hence induce the same effects. One victim has strong family support and lives an unpressurized life, while the other victim lives a poverty-stricken life by himself, having economic or other pressures. The first one survived, while unfortunately, the other one chose to commit suicide.

In Case Nine, several unique characteristics are present. First, the effect mechanism is significantly different from the other case. The EO or cause uses violent language, but with no force to insistently intimidate the affected objects. Second, the differences of EXOs dominated the overall effect. For two victims, though the EO and its effect are the same, the overall effect is totally different, with strikingly different consequences, as shown in Figure 11.11.

11.10 An Education Example: Evaluating a Learning Intervention

Case Ten [4] concerns the evaluation of an educational intervention, such as a new digital learning platform or pedagogical method, aimed at improving students' learning outcomes. In this case, the EO refers to the educational intervention itself: a particular curriculum design, AI-assisted tutoring system, or instructional strategy introduced to enhance teaching and learning effectiveness.

Here, the AO is the individual learner, the minimal observable system in which the overall effect is measured. However, as in previous cases, the observed evaluation outcome, or the overall effect, is not determined by the EO and AO alone. The EXO includes a wide range of contextual factors such as teacher expertise, classroom environment, socioeconomic background, parental support, institutional resources, and even cultural norms surrounding education. These EXOs interact with both the learning process and its outcomes, often confounding attempts to isolate the true effect of the EO.

Importantly, the EO, AO, and EXO in this case are not static entities. Instead, they form an evolving system that unfolds over time and can naturally be described as a time series, a sequence of observations reflecting the dynamics of learning. In this system, the EO, representing a particular educational intervention, exerts its influence directly on the AO, which is the individual learner. The AO "receives" this influence through modifications in its learning process, for example, changes in cognitive engagement, interaction patterns, or responsiveness to feedback, which then propagate to observable learning outcomes such as comprehension, retention, or transfer. In other words, the measured effect emerges from the interaction between EO, EXO, and AO, rather than from either component in isolation.

11.11 Adversarial Exercise Evaluation: A Case of Systemic Confrontation

In a systemic confrontation scenario (Case Eleven [5]), such as a defensive exercise between a defending force and an attacking force, the fundamental question of evaluation is not merely to determine who wins or loses. Most importantly of all, it is to uncover the effect of the EO, e.g., a defensive command and control system, from the overall outcome of the exercise.

The challenge lies in that the observed outcome (e.g., the number of successfully intercepted attacks, response latency, or system survivability) is the result of multiple intertwined influences, not limited to the EO alone. Thus, to perform a scientifically

[4] Dr. Wanling Gao authored this section.
[5] Dr Fanda Fan authored this section.

valid evaluation, one must build an SES that captures all relevant interacting objects and allows the attribution of the true effect of the EO.

In such a confrontation scenario, different from the other cases, the SES embodies a dynamic and interdependent environment: each object not only responds to the others but may also alter its own subsequent states. Therefore, the evaluation must not treat the effect as a static or linear response but as a systemic phenomenon that emerges from the mutual interactions among EO, AO, and EXO in time series.

Also, unlike static laboratory experiments, the SES in a confrontation exercise is *dynamic, adaptive, and reciprocal*: every decision made by the EO induces a counter-reaction from the EXOs, which in turn modifies the operational state of the AO. This cyclical causation challenges the traditional linear model of cause–and–effect analysis and highlights the necessity of viewing evaluation as a *systemic causal inference process* rather than a unidirectional measurement.

The adversarial confrontation scenario provides a rich and realistic context in which the fundamental principles of Evaluatology can be examined and extended.

In adversarial or competitive systems, the role of Evaluatology thus expands from verifying outcomes to uncovering the *structure of influence* that generates them. Through explicit modeling of EO–AO–EXO interactions, evaluators can isolate the true effect of a system even under continuous adaptation and opposition, thereby fulfilling the ultimate goal of Evaluatology: to scientifically uncover the effects of causes within a self-contained, yet dynamically interacting, world.

11.12 Another Physics Example: Verification of a Theory or Model

Case Twelve involves verifying a theory of a model of surface physics, which could be extended to any discipline. Dr. Michael Scriven [175] exemplified this case as the fundamental role of evaluation science, as it is witnessed in almost all disciplines. However, I do not intend to consider this as an evaluation. Instead, it should be categorized into testing, as the aim of this practice is to confirm whether the theory or model conforms to the test oracle. Please see our discussion in 5.8 in Chapter 5.

11.13 Difference from the Other Definitions of Evaluation

In Chapters 1 and 2, I have summarized the unique differences of Evaluatology concepts from those of other works.

In this section, I repeat two points. First, I define the evaluation as uncovering an EO's effects and derived effects, contrasted with determining the merit, worth, or value in [172, 174, 175, 176, 55, 68, 85, 196, 195].

Second, I propose a simple and concise concept system. Instead, the others relied on the encyclopedic approach to define several hundred concepts or terms in [174, 118].

11.14 Summary

This chapter demonstrated that it is feasible to propose a universal and concise concept framework that can be applied across various evaluation scenarios in different disciplines.

Chapter 12

Categories of Evaluation Problems

Per discussion in Section 8.2, the essence of evaluation is to *uncover an EO's effects and derived effects*.

This chapter analyzes the structure of different evaluation problems, which could be based on the category theory introduced in Section 3.7. First, I formalize the evaluation problem and its dual problem, and the inverse problem. Then I discuss the categories of evaluation problems

12.1 Formalization of Evaluation Problem and Its Dual and Inverse Problems

In this section, I formally define the evaluation problem, then introduce the dual problem of evaluation, that is, design, and the inverse problem of evaluation, that is, de-evaluation. Figure 12.1 [1] shows the relationships among evaluation, design, and de-evaluation problems.

12.1.1 Formalization of Evaluation Problem

In this section, I formulate the evaluation problem [2]. I note an EO as O and its element as o. O has n (n is a constant) components F_i, noted as $O = F_1 \times \cdots \times F_i \times \cdots \times F_n$. Each component F_i has $\sigma_i \geq 1$ configurations, where σ_i is a constant. I note the element of F_i as f_{ij_i}, where $j_i \geq 1$ and $j_i \leq \sigma_i$. $o = (f_{1j_1}, \cdots, f_{ij_i}, \cdots, f_{nj_n})$.

I note an EC as C and its element as c. C has m (m is a constant) components G_i, noted as $C = G_1 \times \cdots \times G_i \times \cdots \times G_m$. Each component G_i has $\gamma_i \geq 1$ configurations, where γ_i is a constant. I note the element of G_i as g_{ij_i}, where $j_i \geq 1$ and $j_i \leq \gamma_i$.

[1] Mr. Chenxi Wang contributed to this figure based on the discussion with me.

[2] Mr. Chenxi Wang and Mr. Hongxiao Li also contributed to this section.

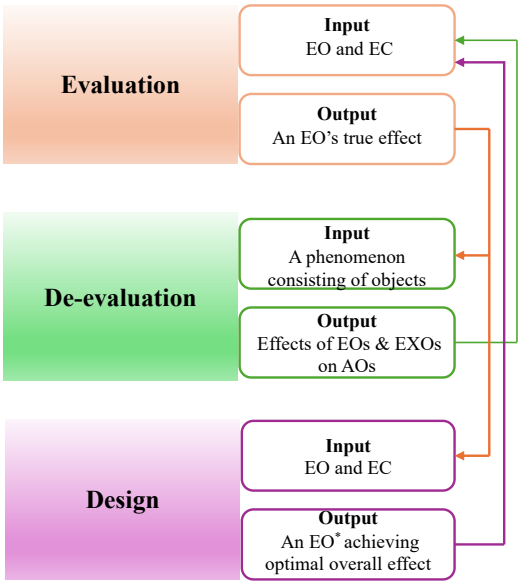


Figure 12.1: The relationship among evaluation, de-evaluation, and design problems.

$c = (g_{1j_1}, \cdots, g_{ij_i}, \cdots, g_{mj_m})$. We note the distribution of a specific configuration of c as $h_{EC}(c)$ where $\sum h_{EC}(c) = 1$.

I note that the overall effect observed on the AO is oe, and the true effect of EO is to be inferred, noted as e. In this context, attributing the overall effect to an EO is formulated as follows.

For a specific EO configuration o, given the overall effect, $oe(o|c)$, under the EC where c follows a distribution $h_{EC}(c)$, how to infer the true effect of the EO, $e(o)$?

12.1.2 Design: The Dual Problem of Evaluation

In this section, I explain why a design problem is the dual problem of an evaluation [3].

Evaluation and design are dual problems. The overall effect is induced by both EO and EXOs on the AO. The evaluation is to uncover an EO's true effect from the overall effect, while the design of an EO aims to search for a specific EO configuration to achieve the optimal overall effect.

12.1.3 De-evaluation: The Inverse Problem of Evaluation

Based on the discussions in Chapter 8, we can easily define the inverse problem of evaluation, which we call de-evaluation.

[3]This idea is inspired by my colleague, Dr. Chunjie Luo, in a group meeting.

I take the diagnosis of traffic jams as an example to showcase a de-evaluation problem. Traffic jams are widely found in many cities. But the traffic jam is a composite phenomenon that reflects the quantities of many objects. The purpose of the diagnosis of traffic jams is to attribute the composite phenomenon to the effects of EOs and EXOs on AOs. In nature, this essential behind the phenomenon is the inverse problem of the evaluation, that is, *how to accurately attribute the effects to their causes*.

Let me explain the inverse problem of the evaluation according to the definition of cause and effect in Section 8.1. We observe a phenomenon P that consists of many objects, O_i, showing different quantities, Q_{ij}, and the changes of quantities, $\mathrm{d}Q_{ij}$ that can be measured or tested. The inverse problem of the evaluation is to trace back the Q_{ij}, and the changes of quantities, $\mathrm{d}Q_{ij}$ of objects, O_i, to the EOs (causes), EXOs, and their effects on the AOs.

12.2 Categories of Evaluation Problems

This section categorizes evaluation problems from different perspectives.

12.2.1 Categories of Evaluation Problems According to the Nature of EOs

According to Section 3.1, an object is a class of entities owning a set of *properties*. Every object could be an EO, so it is also reasonable to classify the evaluation problems according to the nature of the EO.

The first kind of evaluation problem is when the EO is a natural object. I classify it into several sub-categories. The first sub-category is that its properties can not yet be interrogated by a subject. The Second sub-category is that its properties are partially known by the subject. The third sub-category is that the EO can be well-defined, that is, it can be replicated accurately. The fourth sub-category is that the EO can be well-defined, and its components can be freely modified at the subject's discretion.

The second kind of evaluation problem is when the EO is an artifact. According to the definition 3.3, an *artifact* refers to an object that demonstrates intentional conjecture, design, and/or fabrication by a subject. An artifact could be an idea, an imagination of something that does not exist, fake news, or a new design and implementation.

As they are created by a subject, which could be an automatic object or an intelligent life, the unique characteristics of this kind of problem depend on how the artifact induces the effect on the subject, which could be significantly different from that of a natural object.

12.2.2 Categories of Evaluation Problems According to the SESes

According to the definition in Sections 3.5, 10.1, a *model* is a streamlined representation of an object or several objects with mutual effects [221, 224]. I formally call a streamlined representation of an evaluation system *an evaluation model*. Essentially, an SES or a derived SES is an evaluation model. So it is reasonable to classify different evaluation problems according to the nature of the SES.

The first kind of evaluation problem is when an SES is unknown. For example, when we evaluate the parallel universes or the soul or dark matter to reveal their effect, we do not even know what the AOs are.

The second kind is when an SES is only partially known, e.g., when evaluating an object in cosmology and astronomy. The exact number of AOs remains unknown, and it is unclear whether other EXOs exist.

The third kind is when an SES is known. In this case, there are three different subcategories. The first subcategory is when an SES is very complex and cannot be *well-defined*. Being not well-defined indicates that we can not replicate the SES accurately. The second subcategory is when an SES is known and well-defined but not subject to arbitrary manipulation for different reasons, such as realization limitations, unaffordable costs, unaffordable consequences, or ethical reasons. "If a system can be modeled in a function, arbitrary manipulation entails setting its independent variables to any arbitrary number within its domain [222]." The third subcategory is when an SES is known, well-defined, and subject to arbitrary manipulation. For example, a computer nearly falls into this category [222].

12.2.3 Categories of Evaluation Problems According to the Effect Mechanisms

According to the definition in Section 8.1, the effect mechanism is the way through which the cause object induces the effect on the affected objects. Another perspective on analyzing categories of evaluation problems could be through their effect mechanism.

The first dimension is the relationship between the EO and the AO. There is a class of evaluation problems whose common characteristics are worth considering; the EO is a component of the AO. For example, in Case One discussed in Section 11.1, the EO is a CPU, which is the component of the direct AO, a computer system upon which the CPU has an effect.

In Section 11.10, the relationships between the EOs and AOs are much more complex. The genes (components), that is, the EO, of the parent peas (AOs) induce different effects on the descendant peas (AOs). While in the other case, the EO is independent of the AOs. For example, when an EO is a drug, it will affect a specific patient or a population of patients. The EO is independent from the AO.

The second dimension is different effect mechanisms. For example, the Earth could induce gravity on an apple. A CPU works together with several indispensable components and culminates as a computer system; the effect mechanism is the collaboration. A criminal suspect could hurt a victim through intimidation, through insulting or mean language.

The third dimension is whether the interactions of EOs or AOs will impact the effect or not. This case is often found in the policy evaluation, as discussed in Section 11.6. For example, in evaluating a policy to curb drug addiction, the interactions among the AOs, especially their attitude, will impact the evaluation outcome significantly. In the cases discussed in Section 11.10, the interactions of EOs or AOs also impact the evaluation outcome significantly.

12.3 Summary

This chapter formally defines the evaluation problem and its inverse problem and dual problem, while categorizing different types of evaluation problems.

Chapter 13

Fundamental Issues in Evaluatology

In this chapter, I present four fundamental issues in Evaluatology.

13.1 What Evaluation Problems Yield a True or Undefined Evaluation Outcome?

This is the most fundamental issue in Evaluatology. For any evaluation problem that yields an undefined evaluation outcome, we can not test the evaluation approach against the test oracle. Unfortunately, this fundamental issue was never discussed until I formally formulated it in [224].

I address this issue in the proposed evaluation axiom in 9.2. I claimed that when an SES is known, the effect of the EO has the true outcome, which indirectly states what evaluation problem yields a true or undefined evaluation outcome. Other fundamental evaluation approaches, which we will discuss in Part IV, did not explicitly consider this issue.

13.2 Meta-evaluation: Which Evaluation Methodology Yields a Valid Outcome?

As the name of meta-evaluation implies, the purpose of meta-evaluation is to evaluate different evaluation approaches. As I discussed in Chapter 8, for this task, a specific evaluation approach is the EO; the direct AOs are the *inferred effects* of different kinds of EOs that the specific evaluation approach targets. The aim is to reveal the effect of a specific evaluation approach on the inferred effects of different kinds of EOs that the specific evaluation approach targets.

According to the discussions in Chapter 12, how to classify different evaluation problems depends on the nature of the EOs, SESes, and effect mechanisms. I can roughly classify the evaluation problem into two categories of sub-problems.

First, when an evaluation problem has a true outcome, which evaluation methodology can yield the inferred effect approaching the true outcome?

Second, when an evaluation problem has an undefined evaluation outcome, are there any valid evaluation methodologies?

These two fundamental meta-evaluation problems are rarely discussed.

13.3 What Are The types and Natures of EOs, AOs, and Their Effect Mechanisms?

As discussed in Section 8, the essence of the evaluation is to uncover the effect of the EO on the AO. A fundamental question remains: What are the types and nature of the EOs, AOs, and their effect mechanisms?

This is not a trivial issue, as how to uncover the effect of an object is the fundamental issue in almost any discipline. Through a systematic understanding of different EOs, AOs, and their effect mechanisms, we can gain a deeper understanding of the evaluation problems.

13.4 How to Propose Effective and Efficient Evaluation Models?

This section is based on our previous work [224]. I have made significant simplifications.

Effectiveness in evaluation means achieving the desired evaluation outcomes that closely align with the true outcomes, while efficiency refers to achieving these outcomes with minimal resource expenditure.

The key to the effectiveness and efficiency of evaluations in different scenarios is to establish a series of evaluation models that ensure the transitivity of the primary characteristics. I define *a perfect evaluation model* as one under which we can infer the true effect of the EO, e_t.

The *discrepancy threshold* ϵ is defined as *the difference between the true effect* e_t *and the inferred effect* e_i . In reality, how to perform an efficient and effective evaluation with controlled discrepancies is the most important engineering issue. That is how to strike a balance between ensuring the discrepancy threshold and managing the associated costs.

By disregarding the accuracy of an evaluation model, conducting evaluations solely through fully exploring the evaluation model space may indeed result in maximum confidence. However, this approach also comes with a significant drawback - the exorbitant cost it entails.

When creating an evaluation model, the discrepancy threshold ϵ, which is a discrepancy limit that can be tolerated in an evaluation scenario, holds the potential to exert a

profound influence on the evaluation outcome and, in certain instances, it would give rise to grave concerns, particularly in the context of safety-critical tasks where failure could lead to detrimental side effects such as harm, loss of life, or significant environmental damage. So, after thoroughly understanding the stakeholders' evaluation requirements, a risk function $\gamma(\cdot)$ could be predefined. When the stakes are high and there is a greater risk associated with the evaluation outcomes, it becomes imperative to have a lower discrepancy threshold between the inferred effect under an evaluation model and the true effect. This is because the potential consequences of making a wrong decision or drawing inaccurate conclusions become more significant.

In creating an evaluation model, we use the notation $m(\cdot)$ to represent this modeling process. The accuracy of an evaluation model is decided by $m(\cdot)$.

The discrepancy function of the evaluation outcomes $\mathrm{disc}(\cdot)$ between M_g under a specific evaluation model, g, and M_p under a perfect evaluation model, p, is defined as follows:

$$\begin{cases} \text{discrepancy threshold } \epsilon = |\mathrm{e_t} - \mathrm{e_i}|, \\ \text{discrepancy threshold } \epsilon = \gamma^{-1}(\cdot), \\ \mathrm{M_p} = \mathrm{M_p}(\mathrm{k_1}, \cdots, \mathrm{k_n}), \\ \mathrm{disc}(\mathrm{M_g}, \mathrm{M_p}) = \mathrm{disc}\Big(\rho\big(\mathrm{m}(\mathrm{M_p})\big), \rho(\mathrm{M_p})\Big), \\ \mathrm{accuracy}(\mathrm{M_g}) = \mathrm{accuracy}\big(\mathrm{m}(\mathrm{M_p})\big), \\ \mathrm{cost} = \psi(\|\mathrm{M_g}\|). \end{cases} \tag{13.1}$$

In the formulation, $\rho(\cdot)$ is a measurement function. Besides, we define the evaluation cost as the product of a constant ψ and the space capacity of M_g. This cost factor allows us to incorporate the resource constraints and practical considerations associated with the evaluation process. Based on the above formulation, the evaluation issue of balancing evaluation cost and the discrepancies in the evaluation outcomes can be framed as an optimization problem. The objective is to minimize the evaluation cost, represented by cost, while ensuring the discrepancies in the evaluation outcomes, denoted as $\mathrm{disc}(M_g, M_p)$, do not exceed a predefined discrepancy threshold ϵ.

The optimization problem can be formulated as follows:

$$\arg\min \mathrm{cost}(M_g) \quad \text{subject to} \quad \mathrm{disc}(M_g, M_p) < \epsilon, \tag{13.2}$$

or

$$\arg\max \mathrm{accuracy}(M_g) \quad \text{subject to} \quad \mathrm{cost}(M_g) < \varepsilon. \tag{13.3}$$

Another related issue is how to ensure evaluation traceability, which involves attributing any divergence in evaluation outcomes to disparities in the underlying ECs, thereby establishing clear and transparent traceability.

According to the third axiom of evaluation, for the same (derived) EO, the divergence in the (derived) Effect can be attributed to disparities in (derived) ECs, thereby establishing evaluation traceability, establishing evaluation traceability.

Conceptually, traceability asks for a quantified mapping between the differences in the input and output of the measurement function $\rho(\cdot)$ through the evaluation process, described by the mathematical model we formulated above. Our previous work [224] reveals that this concept aligns well with the mathematical notation of the gradients of a function, which gives the rate of change in the output for each input variable.

In the context of evaluation, the gradient of evaluation outcomes can be written as follows, which is a matrix or tensor:

$$\nabla\rho(m(M_p)) = \nabla\rho\Big(m\big(M_p(k_1, \cdots, k_n)\big)\Big) = (\frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_1}, \cdots, \frac{\partial\rho}{\partial m}\frac{\partial m}{\partial M_p}\frac{\partial M_p}{\partial k_n}). \quad (13.4)$$

"∇" is the gradient operator, which in this context represents the vector of partial derivatives of a scalar function with respect to its variables. "∂" is the partial derivative operator, indicating the rate of change of a function with respect to one of its variables while holding others constant.

The closed-form mathematical expression is not always available for various EC components in evaluation. Nevertheless, we can follow the method of acquiring gradients in numerical methods by creating perturbations in the ECs for various input variables and observing the differences in the composite evaluation outcomes, thus approximating the gradients.

13.5 Summary

This chapter presented four fundamental evaluation issues regarding true or undefined evaluation outcomes, meta-evaluation, types and natures of EO, AO, and effect mechanisms, and effective and efficient evaluation models.

Chapter 14

Fundamental Evaluatology Methodology

In this chapter, I present the fundamental Evaluatology methodology.

14.1 Challenges in Real-world Evaluations

We refer to the entire population of real-world systems that are used to evaluate a specific EO as the *real-world evaluation system*. Assuming no safety concerns are present, the real-world evaluation system seems like a reasonable choice for the prime candidate for evaluating the EO. Unfortunately, there are three significant obstacles to consider when assessing diverse EOs with a real-world evaluation system.

Firstly, it is formidable or impossible to isolate an SES from the real-world evaluation system. It is very difficult to recognize EOs, AOs, and EXOs. Also, the presence of irrelevant objects in the real-world evaluation system poses a considerable challenge. It is often difficult, if not impossible, to eliminate these confounders. They can complicate the evaluation process by introducing factors that make it challenging to isolate the effects of other irrelevant objects. Also,

Secondly, manipulating the real-world evaluation system is a formidable task, making it virtually impossible to establish controlled environments for evaluating objects. Even if it is possible to isolate an SES from a real-world evaluation, it is very formidable or impossible to change the state space of the EO, AO, and EXOs. Additionally, the interconnected nature of the components of SES and other irrelevant objects further complicates this challenge.

Thirdly, regardless of the nature of the evaluation problems, there is a tendency for the EOs, AOs, and EXOs to exhibit bias towards certain object instances or their state subspace. This bias towards specific groups can limit our ability to fully explore and understand the entire range of possibilities available to us.

In this context, it is very challenging to propose an effective and efficient evaluation model to address the above issue.

14.2 Universal Evaluation Methodology in Complex Scenarios

In reality, there are evaluation scenarios with different levels of complexity.

For a specific evaluation problem that has the true evaluation outcome, I define the *accuracy* as the absolute value of the difference between the true effect and the inferred effect under an evaluation model divided by the absolute value of the true effect. A lower ratio, closer to 0%, indicates a better accuracy.

For clarity, it is important to distinguish among the *overall effect*, *inferred effect*, and *true effect*. The overall effect refers to the observed evaluation outcome obtained directly from the direct AO, encompassing the joint influence of the EO and EXOs. From this overall effect, the evaluator derives the inferred effect—an estimation of the EO's true effect.

However, the inferred effect does not necessarily equal the true effect. In practice, due to incomplete control or imperfect knowledge of the system, the inferred effect serves only as an approximation of the true effect, and the accuracy metric quantifies the deviation between them.

14.2.1 Concepts of Perfect and Imperfect SESes

I defined *a perfect SES* as *an evaluation model that can fully and accurately identify the EXO that impacts the AO, isolate the irrelevant objects, and infer the true effect of an EO.*

A perfect SES has four unique characteristics that differ from those of real-world evaluation systems. First, it can correctly recognize the EO, AOs, and EXOs. Second, it can completely isolate irrelevant objects. Third, under a perfect SES, we can infer the true effect of the EO. Fourth, we can freely manipulate the full space of SES. This flexibility would enable researchers to explore various scenarios and assess EOs under a range of conditions, enhancing the depth and breadth of the evaluation process.

However, in reality, it's challenging or even impossible to isolate all irrelevant objects or identify all EXO to establish a perfect SES under which we can infer the true effect. Even in the worst cases, we only partially know the SES, as we discussed in Section 14.

I define *the imperfect SES* as *an evaluation model that partially and inaccurately recognizes AOs, EXOs, and isolates the irrelevant objects.* Under an imperfect SES, the true effect of an EO can only be approximately achieved.

While we can strive to create more accurate evaluation models, it is crucial to recognize the inherent limitations and constraints that exist. Due to the limits of our capacity and capability in understanding objects, in reality, it is much more possible that we can only establish an imperfect SES than a perfect SES.

14.2.2 Simple SESes

A perfect or imperfect SES encompasses a vast number of objects, their respective quantities, as well as a large population of EOs, AOs, and EXOs, which leads to significant evaluation costs. However, to address this challenge, it is important to simplify the perfect or imperfect SES in two key ways.

Firstly, to reduce the evaluation costs associated with a large number of objects and their respective quantities, it is crucial to identify and focus on the objects and their quantities that have a significant impact on the evaluation outcomes.

By identifying and prioritizing these targets, researchers can streamline the evaluation models and allocate resources more efficiently. Negligible objects and their quantities, which have a minimal effect, can be excluded or controlled for, thereby reducing complexity and costs. It is worth emphasizing that the simplification will inevitably lead to a decrease in the accuracy of the evaluation model.

Secondly, sampling techniques can be employed to manage the extensive populations of the EO, AO, and EXO. Rather than evaluating every single possibility, researchers can select representative samples that capture the diversity and range of the population. This approach allows for a more manageable evaluation model while still maintaining a good level of coverage and representation.

After these steps, we obtain *a simple SES*, which we define as *the evaluation model that simplifies the perfect or imperfect SES*. A simple SES reduces the objects and their quantities that have little impact on the effect and samples the state space of the perfect or imperfect SES. Accordingly, a simple EC is the simple SES upon the removal of the EO.

A simple SES can be considered a sample of a perfect or imperfect SES. In order to measure the extent to which the statistics of a simple SES can infer the parameters of a perfect SES, we employ the criterion of confidence level and interval.

The confidence level provides us with the probability that the estimated parameters of a perfect or imperfect SES fall within a specific range of values. Meanwhile, the confidence intervals establish a range of values within which we can reasonably expect the true parameters of a perfect or imperfect SES to fall. By utilizing these statistical measures, we can assess the degree of alignment between the statistics of a simple SES and the parameters of a perfect or imperfect SES. This allows us to gauge the efficiency and effectiveness of the simple SES.

By implementing these simplifications in a simple SES, researchers can strike a balance between comprehensiveness and feasibility. The simple SES allows for a more effective and efficient evaluation of EO within the EC framework, mitigating the challenges posed by evaluation costs and the complexity of the perfect or imperfect SES.

14.2.3 Mathematical Formulations of Different SESes

This subsection presents the formulations of different SESes.

First, a sample of the perfect or imperfect SES, M_{ps}, is taken from the perfect or imperfect SES, M_p, and I introduce the notation $s(\cdot)$ to represent the sampling process. Additionally, the Simple SES M_s reduces the independent variables within the sample of the perfect or imperfect SES, M_{ps}, by excluding those that have minimal impact. To formalize this process, I introduce the notation $m(\cdot)$ to represent the modeling process consistent with that in Section 13.4.

When transforming a perfect or imperfect SES into a simple SES, it is essential to maintain the transitivity of the following characteristics:

$$M_{ps} = s(M_p). \tag{14.1}$$

The sample of the perfect or imperfect SES M_{ps} is obtained through a sampling process $s(\cdot)$ on the perfect or imperfect SES M_p.

$$M_{ps} \subset M_p. \tag{14.2}$$

The sample of the perfect or imperfect SES M_{ps} is a subset of the perfect or imperfect SES M_p.

$$M_s = m(M_{ps}). \tag{14.3}$$

The simple SES M_s is obtained through a modeling process $m(\cdot)$ on the sample of the perfect or imperfect SES, M_{ps}.

For each element m_s in the simple SES M_s, we denote the corresponding element in M_{ps} as $m_{p's} \in M_{ps}$.

$$m_s = (k_1, \cdots, k_{n'}) \in M_s. \tag{14.4}$$

Each element m_s in the simple SES M_s consists of a set of independent variables $(k_1, \cdots, k_{n'})$.

$$m_{p's} = (k_1, \cdots, k_{n''}) \in M_p, \text{ where } n' \leq n''. \tag{14.5}$$

The corresponding element $m_{p's}$ in the sample of the perfect or imperfect SES M_{ps} consists of a set of independent variables $(k_1, \cdots, k_{n''})$. The n'' is greater than n', ensuring that the corresponding element in the sample of the perfect or imperfect SES includes at least as many independent variables as that of the element in the simple SES.

14.3 What is a Benchmark?

Benchmarks are extensively employed across various disciplines, albeit lacking a formal definition. Based on the science of evaluation, we propose a precise delineation of a benchmark as *the EXOs in a simple EC* according to the definition in Sections 14.2.1 [1].

[1] This definition is based on the discussion with Dr. Lei Wang and Mr. Chenxi Wang.

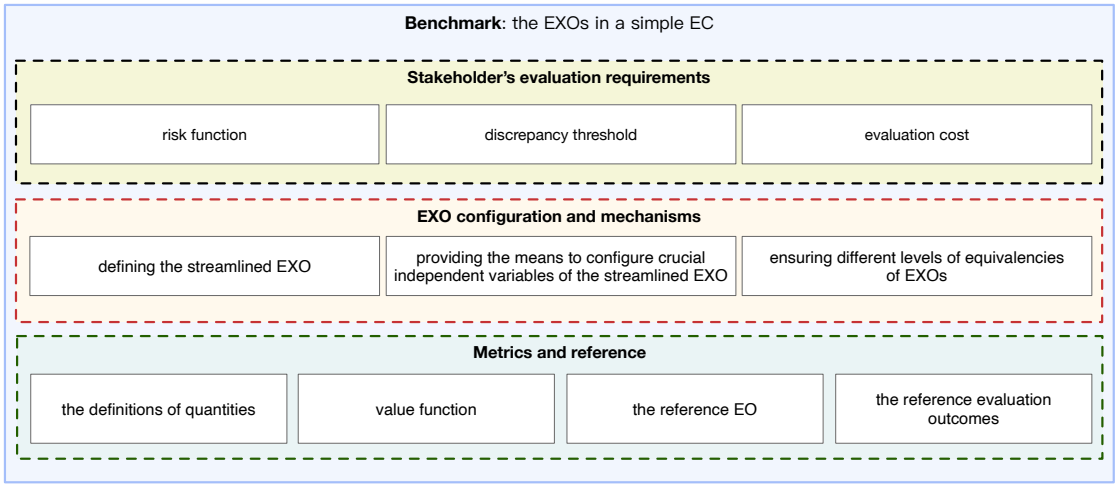


Figure 14.1: A benchmark comprises three essential constituents.

Specifically, a benchmark refers to the EXOs in a simple EC, where the combined effect of the EXOs and the EO is exerted on the AO.

Within the framework of this definition, a benchmark comprises three essential constituents, as shown in Figure 14.1. The first constituent is the *stakeholder's evaluation requirements*, which encompass various factors. These include the risk function, which evaluates the potential risks associated with the benchmark. Additionally, the discrepancy threshold, which determines the acceptable level of deviation from the true evaluation outcomes, is considered. Lastly, the evaluation cost is taken into account, and the resources required for conducting the evaluation are assessed. By considering these elements, the benchmark can effectively address the evaluation requirements of stakeholders.

The second constituent of the benchmark framework is the *EXO configuration and mechanisms*. This includes several elements crucial for the benchmark's effectiveness. It involves defining the streamlined EXOs; it provides the means to configure crucial independent variables of the streamlined EXOs; it ensures different levels of equivalencies of EXOs. By considering these EXO configurations and mechanisms, the benchmark can provide a comprehensive and standardized approach to different evaluation issues.

The third constituent is the *metrics and reference*, which includes the definitions of quantities, the value function, the reference EO, and the reference evaluation outcomes.

14.4 Fundamental Methodology Under a Perfect SES

As shown in Figure 14.2, this section proposes a fundamental Evaluatology methodology under ideal conditions where we can isolate and manipulate a perfect SES [2].

[2]Dr. Wanling Gao, Mr. Chenxi Wang, Dr. Lei Wang, and I contributed to this section.

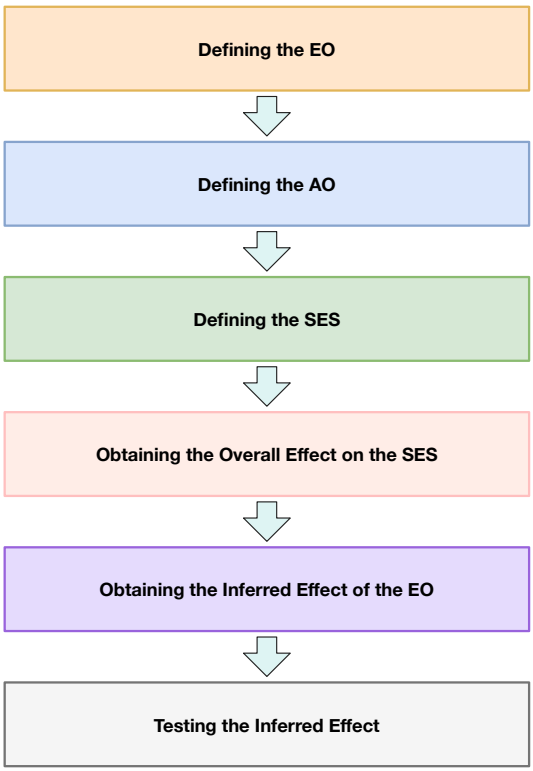


Figure 14.2: A fundamental Evaluatology methodology under ideal conditions where we can isolate and manipulate a perfect SES.

14.4.1 Defining the EO

The first and most fundamental step in Evaluatology is to clearly characterize and define the EO. This process goes far beyond merely naming or labeling the EO; this requires rigorous specification of its boundaries and configurations, including the internal components and their interconnections. Without such a precise definition, subsequent comparisons across different EOs or even across different instances of what is assumed to be the "same" object are rendered ambiguous and potentially invalid. The EO is the *cause object* in the causal structure of evaluation—it is the entity whose effect we aim to uncover or quantify.

A central motivation for this step lies in the principle of comparability. In any evaluation, the validity of results depends on ensuring that the EOs under study are indeed instances of the same class according to the definition of object in Section 3.1. For example, two entities may both be labeled as "processors", but if one refers to a central processing unit (CPU) and the other to a graphics processing unit (GPU), then their fundamental architectures, functions, and operational contexts differ to such an extent that direct comparison would yield misleading or nonsensical conclusions. Even within a narrower category, such as CPUs, the EO may not be monolithic: the same processor can operate under multiple configurations (e.g., with or without turbo boost enabled, or under different thermal or power management settings). Without explicitly specifying which configuration is being treated as the EO, the evaluation outcomes may vary widely and unpredictably.

Following the notation defined in Chapter 12 Formally, we can represent an element of the configuration space of a well-defined EO as: $o_i = \{o_{i1}, o_{i2}, ..., o_{ix}, ...\}$. Here, i denotes the instance of the EO (e.g., "CPU model X"), while x enumerates the possible configurations (e.g., "default mode", "turbo boost enabled", "low-power mode", etc.). Each element o_{ix} thus corresponds to a specific operational configuration of the i_{th} component. Selecting which configurations are to be evaluated is not merely a technical choice but a scientific one, since it determines the interpretability and comparability of the evaluation results.

From a practical perspective, the configuration space O_i is often extremely large, and it is usually neither feasible nor necessary to evaluate the object under every possible configuration. Consequently, evaluators often reduce this space by selecting a representative configuration o_{ij}. Such a configuration is typically chosen because it achieves high or stable performance under a broad set of ECs, or because it reflects the most common real-world usage scenario. For example, in CPU evaluation, one might fix the processor to its default settings or select the configuration that yields the best performance on a majority of standard workloads.

However, this strategy comes with a critical caveat: no single configuration is guaranteed to perform optimally across all ECs. A CPU setting that excels in compute-intensive workloads may underperform in energy-constrained scenarios. Similarly, a software algorithm tuned for accuracy may sacrifice efficiency when faced with real-time constraints. Therefore, while adopting a representative configuration can greatly reduce the evaluation cost and complexity, it also risks oversimplification, potentially leading to biased

or incomplete conclusions.

This tension—between comprehensiveness (evaluating all possible configurations) and feasibility (reducing the configuration space)—is at the heart of defining the EO. Striking the right balance requires a careful understanding of both the object's intrinsic variability and the intended scope of the evaluation. If the goal is a broad scientific characterization, more configurations may need to be included. If the goal is practical deployment guidance, a smaller number of well-chosen configurations may suffice.

In summary, defining the EO is not a trivial preliminary step but a foundational process that determines the validity and interpretability of the entire evaluation. It ensures that comparisons are meaningful, outcomes are reproducible, and conclusions can be properly attributed to the EO itself rather than to hidden or uncontrolled variations in its configurations.

14.4.2 Defining the AO

Once the EO is clearly defined, the second step in Evaluatology is to specify the AO—the minimal system on which the effects of the EO can be directly measured or tested, as we defined in Section 8.2.

The AO serves as the *effect-receiving system*. The definition of AO is critical, as it determines the scope of measurement or testing.

Formally, the AO can be conceptualized as a composed system:

$$a_j = \{c_{j1}, c_{j2}, ..., c_{jk}\}, \tag{14.6}$$

where c_{jk} represents the k_{th} configuration of j_{th} *indispensable components*—the minimal set of functional elements that collectively allow the EO's effects to manifest and be empirically measured or tested. The AO must be sufficiently complete to support the EO's operation, yet minimal enough to avoid the introduction of unnecessary confounding factors. This principle—*completeness without redundancy*—ensures that every observed effect is causally meaningful and traceable to the EO within the defined boundaries of the system.

In one category of evaluation problems, the EO cannot function independently; it must be embedded within a broader system that provides essential infrastructure for its operation. For example, when the EO is a central processing unit (CPU), it cannot be evaluated in isolation, as it depends on memory, storage, and input/output subsystems to execute any meaningful workload. Here, the AO is the *computer system* itself—a minimal operational environment that includes all indispensable components required to sustain computation. Specifically, such an AO typically consists of:

- *Primary computational unit (CPU):* the EO whose performance or characteristics are under evaluation;
- *Memory subsystem:* provides working storage for instruction and data execution, directly influencing latency and throughput;

- *Persistent storage:* enables data loading, checkpointing, and I/O operations necessary for full program execution;
- *Interconnect and power subsystem:* ensures communication and energy delivery among components, maintaining operational stability.

These indispensable components form the minimal closed system in which the EO's effects (e.g., execution time, power consumption, cache efficiency) can be observed, measured, and meaningfully compared. Adding or omitting any of these components would fundamentally alter the measurement context and, therefore, the interpretation of the results.

In other categories of evaluation problems, the AO does not necessarily have to include the EO itself. In many domains, particularly outside of engineering, the AO may exist as an independent system that responds to the EO's intervention or influence. In pharmacological evaluation, for instance, the EO is the pharmaceutical compound, while the AO is the human body, a complete biological system within which the EO's biochemical effects manifest. The body already constitutes a self-sufficient measurable system, and the EO acts upon it externally. This case contrasts with some engineering systems, where the EO is typically an internal component of the AO.

Thus, depending on the domain and causal structure, the EO–AO relationship can be classified into two general categories:

- *Inclusive Relation:* The AO includes the EO as one of its indispensable internal components, such that the EO operates as a subsystem within a larger measurable whole (e.g., CPU within a computing system, or an engine within a vehicle).
- *External Relation:* The AO exists independently of the EO, but serves as the environment or organism upon which the EO acts (e.g., a human body receiving a drug, or an ecosystem exposed to a pollutant).

Regardless of the relation type, defining the AO requires balancing two fundamental criteria: *operational completeness* and *causal clarity*. An AO that is too narrow will fail to capture the full manifestation of the EO's effect. At the same time, one that is too broad may dilute the EO's contribution with unrelated interactions. For instance, evaluating a CPU within a large-scale data center might introduce network latency, scheduling, and resource-sharing effects unrelated to the CPU's intrinsic behavior. Conversely, evaluating it in a minimal system (with controlled memory and I/O subsystems) provides a clearer mapping from cause (EO) to effect (AO).

In practice, defining a suitable AO involves iterative abstraction and reduction, which depends on the expertise. Evaluators begin by identifying all components required for the EO to function, then systematically remove non-essential elements until a stable and measurable system remains. This minimal AO represents the tightest causal interface through which the EO's effect can be empirically observed without external confounding interference.

In summary, the AO is not merely the recipient of the EO's influence but the operational bridge between the EO's intrinsic behavior and the measurable or testable

outcomes that form the basis of evaluation. A well-defined AO ensures that evaluation results are both *causally valid*—accurately reflecting the EO's true effect—and *empirically grounded*—derived from a physically coherent and reproducible system. Defining the AO thus transforms evaluation from a descriptive exercise into a structured causal analysis, enabling consistent interpretation across domains and systems.

14.4.3 Defining the SES

After defining both the EO and the AO, the next step in Evaluatology is to establish the SES. The SES serves as a self-contained and isolable environment within which the causal effect of the EO can be meaningfully observed, measured, and analyzed. "Self-contained" indicates that the SES incorporates all EXOs that might affect the measurable or testable outcome. It is "isolable" in that other objects can be disentangled from those objects included in the SES.

Formally, we can express the SES as:

$$SES = \{EO, AO, EXO\}, \tag{14.7}$$

where the *EO* represents the causal source under investigation, the *AO* is the measurable system on which the EO's effects manifest, and the *EXO* (Essential External Objects) encompasses all additional objects that can influence the AO alongside the EO.

(1) The structural role of SES. The SES defines the boundary of what is under evaluation. In practice, both EO and EXO exert influence on the AO, and the measured or tested outcome represents the *overall effect*—the joint consequence of EO and EXO on AO. The central objective of evaluation, therefore, is not simply to measure this overall effect, but to *isolate and quantify the true effect of the EO* within it. Only by distinguishing the EO's true effect from that of EXO can one make valid causal claims and ensure that observed differences truly reflect changes in the EO rather than uncontrolled variation in external conditions.

(2) Composition and interactions. The SES captures the structural and causal relationships among the three components:

$$EO \rightarrow AO \leftarrow EXO. \tag{14.8}$$

The arrow from EO to AO represents the causal path of interest—the true effect we seek to uncover—while the influence from EXO to AO represents external interference, modulation, or context. In many cases, EO and EXO may also interact, forming compound effects (e.g., a CPU's performance interacting with the compiler or workload). The design of the SES, therefore, requires careful selection and control of EXO to make EO's effect identifiable.

In biomedical evaluation, if the EO is a pharmaceutical compound and the AO is the human body, the SES also includes EXOs such as environmental conditions, lifestyle factors, or psychological stressors. These EXOs likewise influence the measured outcomes (e.g., blood concentration, physiological response) and must be controlled or randomized to separate the drug's intrinsic effect from external perturbations.

(3) Minimal completeness and causal isolation. The definition of SES follows two guiding principles:

1. *Minimal completeness:* The SES must include all indispensable entities necessary for the EO's effect to manifest on the AO—nothing more, nothing less.
2. *Causal isolability:* The SES must be structured such that the EO's effect can be distinguished, either through experimental control (fixing EXO) or inferential modeling (adjusting for EXO).

In biomedical or social systems, minimal completeness may consist of the EO (e.g., drug, policy), the AO (e.g., patient, population), and EXOs (e.g., environment, behavioral factors). Across both cases, isolating the EO effect requires that EXOs be explicitly identified, documented, and either held constant, randomized, or modeled.

(4) Practical implications. Defining a proper SES is not merely a formal step—it has practical implications for experimental design, reproducibility, and interpretation. An ill-defined SES can lead to confounded results where observed performance differences stem not from the EO but from uncontrolled EXO variations. Conversely, an overly restricted SES may eliminate meaningful variability, leading to results that lack external validity. The art of evaluation lies in finding the correct balance between environmental control and representational realism.

(5) Summary. In summary, the SES represents the minimal evaluation environment that integrates:

- the *EO* —the causal source under investigation,
- the *AO* —the measurable or testable system receiving the effect, and
- the *EXO* —the essential external factors jointly influencing the AO.

The SES establishes the causal stage upon which evaluation occurs: EO and EXO jointly act on AO to produce observable outcomes. Evaluation, in turn, seeks to disentangle these overall effects to reveal the EO's true effect. Without a rigorously defined SES, evaluation becomes ambiguous—its conclusions may reflect uncontrolled context rather than true causality. By grounding evaluation in the structured triplet $\{EO, AO, EXO\}$, Evaluatology transforms evaluation from a descriptive practice into a scientifically interpretable causal analysis.

14.4.4 Obtaining the Overall Effect on the SES

Once the SES is clearly constructed, comprising the EO, the AO, and the EXO, the next stage is to obtain the *overall evaluation outcome* on the SES, which is the overall effect on the AO induced by the EO and the EXO.

This process represents the empirical phase of Evaluatology, where the theoretical structure of the SES is instantiated into measurable experiments or simulations. The overall effect is not the true effect of the EO alone, but the observable result that arises from the combined interaction of EO, AO, and EXO under a wide range of ECs.

14.4.5 Obtain the Inferred Effect of the EO

After obtaining the overall evaluation outcomes on the SES, the next and most critical stage of Evaluatology is to *infer the effect of the EO* on the AO, which we call *the inferred effect.* Please note the difference between the inferred effect and the true effect: the true effect is the target that the inferred effect approaches.

This step transforms the descriptive phase of evaluation—where the overall outcomes are observed—into a causal phase, where the effect of the EO is isolated and quantified. In essence, the goal is to distinguish the portion of the overall effect that is truly attributable to the EO itself, as opposed to the contributions of the EXOs or random variations within the ECs.

14.4.6 Testing the Inferred Effect

After the inferred effect of the EO has been obtained, the final step of Evaluatology is to validate this inferred effect. This stage treats the inferred effect of the EO as a statistical hypothesis and examines whether it is consistent with empirical evidence collected from additional ECs or repeated measurements. The goal is to determine whether the inferred effect genuinely reflects the true effect of the EO, rather than random fluctuations or residual confounding. Please refer to Section 5.9.

14.5 Summary

This chapter presented three evaluation models: perfect, imperfect, and simple SESes, formally defined the benchmark, and proposed a fundamental evaluation methodology under a perfect SES.

Chapter 15

Hierarchical Formalizations of Well-defined SESes

Based on our previous work [224], this chapter presents hierarchical formalizations of well-defined SESes, which work for a class of evaluation problems where EC can be hierarchically defined. This kind of evaluation problem often manifests in the field of computer science.

15.1 Hierarchical Definition of SESes

In Chapter 8, I presented the definition of an SES. There are many ways to define an SES. In this section, we propose a hierarchy definition of an SES. This approach works quite well for a class of evaluation problems for which the EC and the EO can be well-defined. That is to say, we can replicate an SES fully accurately.

We begin by defining an SES from the problems or task spaces that the EO's stakeholders need to address. The reason is as follows. First, the concerns and interests of the relevant stakeholders are at the core of the evaluation. These concerns and interests are best reflected through the problems or tasks they must address, which provide a reliable means to define an SES. Second, utilizing the same problem or task can ensure the comparability of evaluation outcomes.

Taking Case One, presented in Chapter 8, as an example, we observe that in Case One, an SES encompasses numerous constituents. Notably, we identify five primary components of an SES from top to bottom. Figure 15.1 shows a typical example.

The first top component is a set of equivalent definitions of problems or tasks (in short, problems). While the problem itself serves as the foundation for the evaluation process, it cannot solely serve as the evaluation itself because the problem is often abstract and requires further instantiation to determine its specific parameters. The second component is a set of collective problem instances, each of which is instantiated from an element of the first component. Unlike the first component, an equivalent problem instance is specific and can serve as the evaluation directly.

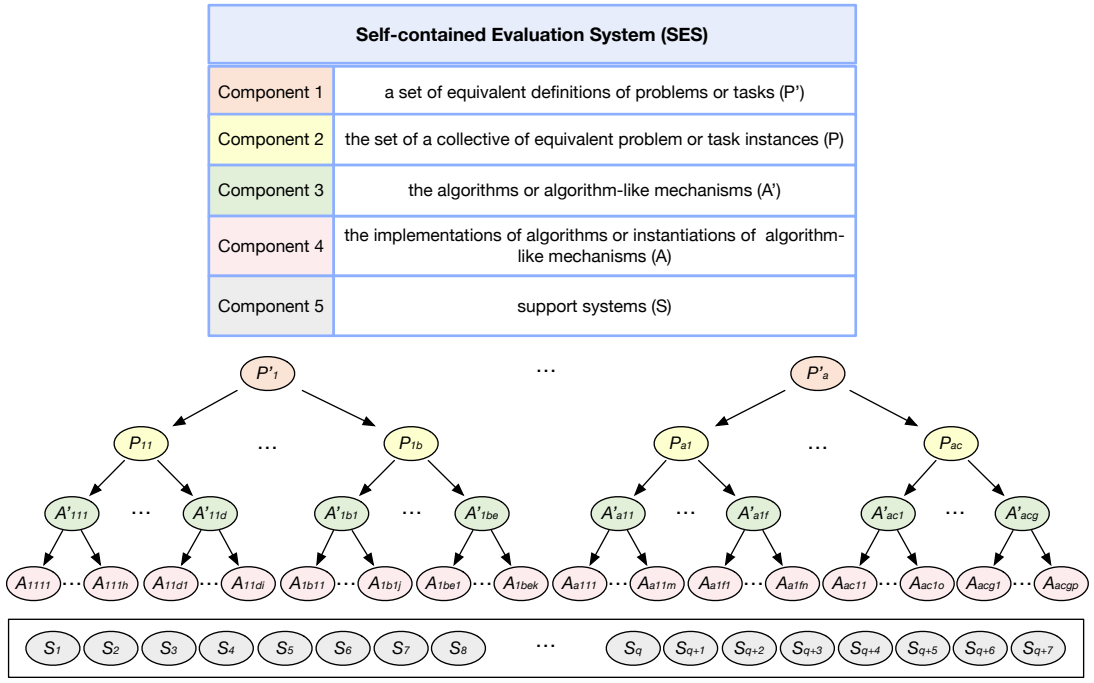


Figure 15.1: The hierarchical definition of an SES.

After a problem instance is proposed, it is necessary to figure out a solution. The third component consists of the algorithm, each of which provides the solution to a specific problem instance.

The fourth component encompasses the implementation of an algorithm. Its implementation involves understanding the algorithm and implementing it in a specific system, which I call *the support system*. This process ensures that the algorithm can effectively and efficiently solve the intended problem instance within the given context.

15.2 Formalization of SESes

To lay the groundwork for the formalization of an SES, it is essential to establish a clear understanding of some crucial notations. The notations P' and S represent two crucial components. P' represents the problem space. S represents the support system space. p', s, o is an element of P', and S, respectively. We note $p' \in P'$, $s \in S$.

In addition to the aforementioned notations, we also define several other fundamental notations. For each problem, $p'_i \in P'$, there is a set of problem instances noted as P_i. For all problems in P', there is an entire collection of problem instances, which can be noted as (p'_i, P_i). The entire set of problem instances P can be defined as the union of all P_i.

We introduce the notation A' to represent the algorithm space. for each problem instance p_{ij} in the problem instance space P_i, the algorithm space A' could be noted as

(p'_i, p_{ij}, A'_{ij}). A' is a union of all A'_{ij}.

We introduce the notation A to represent the instantiations of the algorithm space. This space, A, consists of instantiations of algorithms that are associated with each problem instance, algorithms, and support systems. Specifically, for a given problem p'_i in the problem space P', for each instance p_{ij} in the corresponding instance space P_i, for each algorithm a'_{ijk} in the algorithm space A'_{ij} on each support system s_l in the support system space S, we define the set of instantiations of algorithms as $(p'_i, p_{ij}, a'_{ijk}, s_l, A_{ijkl})$. A is a union of all A_{ijkl}.

By introducing these notations, we establish a comprehensive framework that allows us to delineate the various components of an SES and their respective roles. This formalization enhances our understanding of the key components and their relationships within the EC framework.

We note the EC space as C and the EO space as O. Based on the concepts of the SES defined in Chapter 14, we formalize an SES as $M = C \times O = P' \times P \times A' \times A \times S$.

Likewise, we use symbols to denote a real-world evaluation system (M_r), a perfect or imperfect SES (M_p), and a simple SES (M_s). These are represented as:

$$M_r = P'_r \times P_r \times A'_r \times A_r \times S_r, \tag{15.1}$$

$$M_p = P'_p \times P_p \times A'_p \times A_p \times S_p, \tag{15.2}$$

$$M_s = P'_s \times P_s \times A'_s \times A_s \times S_s. \tag{15.3}$$

These symbols help us distinguish and calculate the various components of ECs and SESes in different contexts.

In the realm of EC spaces, the concept of equivalent ECs plays a significant role. Two EC spaces, denoted as C_1 and C_2, are considered to be equivalent ECs if and only if there exists a bijection, denoted as β, between the two spaces: $\beta : C_1 \mapsto C_2; \beta^{-1} : C_2 \mapsto C_1$. A bijection is a function that is both injective (one-to-one, no two inputs map to the same output) and surjective (onto, every output has a corresponding input), creating a perfect one-to-one correspondence between two sets. This equivalence is denoted as $C_1 \sim C_2$.

15.3 Summary

This chapter presents a hierarchical definition of a well-defined SES that consists of problems, problem instances, algorithms, algorithm implementations, and support systems, which are often found in computer science and engineering. Finally, we formally formalize SESes mathematically.

Part IV

Other Fundamental Evaluation Methodologies

Chapter 16

Design of Experiments

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the Design of Experiments (in short, DoE).

16.1 Basic Concepts

The basic concepts come from early work of DoE [91, 200, 100, 60, 59].

Experimental Units are objects under study, ranging from an individual, e.g., a person, to aggregated entities (e.g., a class).

Responses are quantitative metrics obtained on the experimental units through measurements, distinct from categorical variables.

Factors represent parameters whose changes in value will result in variations in the response. A factor has several levels. The *factor level* could be both the quantitative variable and the categorical variable.

Interactions are any systematic dependencies between factors.

Experimental Condition Groups are the combinations of all factor levels. If there are n input factors, and each factor i has m_i factor levels, then there are a total of $\prod_{i=1}^{n} m_i$ experimental condition groups.

An *Experiment* is to apply different experimental condition groups to the experimental units with determined and controlled factor levels.

Factor Effects are the response difference between the average of experimental condition group runs at the different levels for corresponding factors.

A *Model Equation* is an equation that establishes the model to characterize the relationships between responses and factor effects.

$l^k r$ *Factorial Designs* select k factors, each factor selects l levels, and each experimental condition group replicates r times. Its model equation calculates all effects of selected factors or their interactions.

$2^{k-p} r$ *Factorial Designs* select k factors, each factor selects 2 levels, with 2^p factors or interactions are confounded to reduce costs, and each experimental condition group

[1] Mr. Chenxi Wang is the primary contributor of this author.

replicates r times. Its model equation calculates the aggregated effects from 2^p factors or interactions.

Full Factorial Designs separately estimate all effects of selected factors or their interactions through the model equation. Full factorial designs include $2^k r$, $l^k r$, and general factorial designs. Here, 2 or l indicates the factor level, k means the number of factors, and r is the number of replications. General factorial design also involves studying k factors with r replications, but the selection of levels for each factor is not restricted to the same fixed value (such as 2 in a $2^k r$ factorial design or l in $l^k r$ factorial design).

Fractional Factorial Designs compute aggregated effects of selected factors or their interactions through the model equation. Classical fractional factorial designs include $2^{k-p} r$ factorial designs.

Residual error represents the portion of the response variation not explained by the model that characterizes the relationships between responses and factor effects.

Homoscedasticity means constant variance, while *Heteroscedasticity* indicates non-constant variance.

16.2 Problem Statement

DoE is a structured, statistical methodology developed to systematically infer a linear model between multiple input factors and output responses within a process or system [200, 91, 60]. This approach has become foundational in fields ranging from engineering and manufacturing to biology and behavioral sciences, providing researchers with a powerful tool set for optimization, quality improvement, and hypothesis testing [100, 59, 60].

Given an experiment unit or a population of experiment units with one or more measurable responses of interest and a set of controllable input factors, each factor can be set to different levels. The factor effects, which are the additive components that constitute the expected value of the response, are unknown.

The DoE problem could be formally stated as "how to strategically select a limited but representative set of factor-level combinations to run, allowing for the efficient and simultaneous estimation of the factor effects, to derive a reliable model that describes the relationship between the response and the factor effects, while minimizing the confounding of uncontrolled or unknown factors."

The inputs of DoE include factors, the corresponding levels for each factor, experimental units, and the responses measured from them. The outputs of DoE include the relationship between each response and factor effects. The constraint is that experimental resources are limited; it is infeasible to run all possible combinations of factor levels due to prohibitive costs or time constraints. Furthermore, the observed response is often contaminated by the uncontrolled or unknown factors that may influence the response, which obscures the factor effects.

16.3 Basic Assumptions

Assumption 16.1 (Linearity)**.** *The relationship between the response variable and the factor effects, including their interactions and random error, can be adequately represented by a linear model. Formally, the response Y is a linear function of the factor effects:*

$$Y = \mu + \beta_1 + \beta_2 + \cdots + \beta_{i,j} + \cdots + \epsilon, \tag{16.1}$$

where Y is the response, μ is the overall mean of all responses, β's are the factor effects, and ϵ is the random error term. A single subscripted β denotes the main effect of each factor, with the subscript indicating the corresponding factor; a multi-subscripted β represents the interaction effect among the factors indicated by its subscripts.

Assumption 16.2 (Additivity of Effects)**.** *The effects of different factors are separable and additive in nature. That is, the change in the response caused by varying one factor, to a first approximation, is independent of the levels of other factors. Any interaction can be explicitly identified and modeled as a separate term in the model equation, preserving the principle of effect separation.*

Assumption 16.3 (Randomization)**.** *The order in which experimental runs are performed should be randomized. Randomization serves to distribute the potential effects of uncontrolled or unknown factors that may influence the response evenly across all experimental groups. This process mitigates the risk of confounding these nuisance factors with the systematic effects of the factors under study, thereby ensuring unbiased estimates of the effects.*

Assumption 16.4. *The residual error terms are independently and identically distributed (i.i.d.). Specifically, they follow a normal distribution with a mean of zero and constant variance σ^2:*

$$\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{16.2}$$

This implies independence of errors, homoscedasticity, and normality.

Assumption 16.5. *The variance of the residual errors is constant across all factor levels. Denoted as $Var(\epsilon) = \sigma^2$, this property of homoscedasticity is crucial for the validity of standard hypothesis tests (e.g., F-tests) and the correctness of confidence intervals derived from the model. Heteroscedasticity can invalidate these inferences.*

16.4 Basic Principles

DoE uses a linear model (called the model equation) that sums additive effects from different factors. For each measured response in the experimental units, its effects comprise the overall mean, the main effects of individual factors, interaction effects between factors, and residual random errors—more details in Section 16.5.2.

The Analysis of Variance (ANOVA) methodology decomposes total response variance into components attributable to between-group variance, within-group variance, and the

square of errors, as elaborated in Section 16.5.2. Through ANOVA [167], it quantifies the contribution of each factor to response variance and tests the statistical significance of factors and their interactions.

By systematically combining factor levels into experimental condition groups and applying the experimental condition groups to the experimental units, DoE enables measurement of responses under structured and controlled conditions. It encompasses three effect components: 1) main effects of individual factors, 2) interaction effects between factors, and 3) random errors from chance fluctuations.

Rather than altering one factor at a time, DoE introduces a framework for simultaneously varying multiple factors, thereby enabling the identification of interaction effects and the construction of predictive models. Typically grounded in factorial or fractional factorial designs, DoE incorporates randomization, replication, and blocking to reduce experimental error and enhance the robustness of inferences.

16.5 Methodology

16.5.1 Classical Factorial Design

The factorial design encompasses both full factorial and fractional factorial designs. Full factorial designs separately estimate all main effects and interactions of selected factors through the model equation, while fractional factorial designs compute aggregated effects of multiple factors or their interactions. Although a full factorial design requires higher experimental and computational costs, it provides a precise estimation of individual factor effects and interactions. In contrast, a fractional factorial design reduces experimental costs by estimating aggregated effects, but these are subject to confounding due to the cumulative nature of multiple factor contributions.

- *Full Factorial Designs:* Full factorial designs include $2^k r$, $l^k r$, and general factorial designs. In a $2^k r$ factorial design, each factor selects 2 levels; a $l^k r$ factorial design selects l levels per factor; and a general factorial design allows variable levels across factors. The model equation for full factorial designs accounts for $(2^k - 1)$ total effects, including $\binom{k}{1}$ main effects, and $\sum_{i=2}^{k} \binom{k}{i}$ interaction terms. With r replicates per run, an additional random error term is incorporated. The details are shown in Section 16.5.2. $\binom{k}{i}$ is read as "k choose i."
- *Fractional Factorial Designs:* A fractional factorial design, such as $2^{k-p} r$ factorial design, estimates $(2^{k-p} - 1)$ effects, each representing aggregated effects from 2^p factors or interactions. Similar to a full factorial design, it includes r replicates and a random error term in the model equation.

16.5.2 Model Equations and Analysis of Variance Formulations

We illustrate model equations and analysis of variance (ANOVA) formulations using a general factorial design as a case study. Notably, all other factorial designs are special cases of this approach, sharing the same structural formulation.

Consider a general factorial design investigating k factors influencing the responses. The ith factor has n_i levels, and each experimental run is replicated r times. The model equation decomposes each measurement into a linear additive model comprising: 1) *Overall mean:* The overall average response across all experiment groups; 2) *Main effects:* $\binom{k}{1} = k$ terms representing the individual effect of each factor; 3) *Interaction effects:* $\sum_{i=2}^{k} \binom{k}{i} = 2^k - 1 - k$ terms, including $\binom{k}{2}$ two-factor interactions, $\binom{k}{3}$ three-factor interactions, up to the k-factor interaction; 4) *Random errors:* Accounting for chance fluctuations introduced by repeated measurements.

ANOVA proceeds by squaring and summing all effects, which partition the total variance into distinct components: main effects, interactions, and residual error.

For concreteness, we demonstrate the model equations and the ANOVA formulation with $k = 3$ factors, highlighting how effects are structured and analyzed. Other factorial designs follow an analogous formulation.

In a general factorial design with $k = 3$ factors, Factor A has n_1 levels, Factor B has n_2 levels, and Factor C has n_3 levels. Each experimental combination is replicated r times. The model equation is expressed as:

$$Y_{i,j,k,l} = \mu + a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{16.3}$$

In this model equation, $Y_{i,j,k,l}$ represents the measured response of replicate l at level i of Factor A, level j of Factor B, and level k of Factor C. μ represents the overall mean of all responses. a_i denotes the main effect of Factor A, b_j indicates the main effect of Factor B, and c_k signifies the main effect of Factor C. The two-way interaction effects are captured by $d_{i,j}$ (Factor A × Factor B), $e_{i,k}$ (Factor A × Factor C), and $f_{j,k}$ (Factor B × Factor C), while $g_{i,j,k}$ represents the three-way interaction effect among all three factors. Finally, $\epsilon_{i,j,k,l}$ accounts for the random errors introduced by experimental replication of chance fluctuations.

To calculate these components in the model equation, we first calculate the group means.

$$\overline{Y}_{.,.,.,.} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times n_3 \times r}, \tag{16.4}$$

$$\overline{Y}_{i,.,.,.} = \frac{\sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times n_3 \times r}, \tag{16.5}$$

$$\overline{Y}_{.,j,.,.} = \frac{\sum_{i=1}^{n_1} \sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_3 \times r}, \tag{16.6}$$

$$\overline{Y}_{.,.,k,.} = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times n_2 \times r}, \tag{16.7}$$

$$\overline{Y}_{i,j,.,.} = \frac{\sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_3 \times r}, \tag{16.8}$$

$$\overline{Y}_{i,.,k,.} = \frac{\sum_{j=1}^{n_2} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_2 \times r}, \tag{16.9}$$

$$\overline{Y}_{.,j,k,.} = \frac{\sum_{i=1}^{n_1} \sum_{l=1}^{r} Y_{i,j,k,l}}{n_1 \times r}, \tag{16.10}$$

$$\overline{Y}_{i,j,k,.} = \frac{\sum_{l=1}^{r} Y_{i,j,k,l}}{r}. \tag{16.11}$$

Using these group means, the components of the model equation in Equation 16.3 are calculated as:

$$\mu = \overline{Y}_{.,.,.,.}, \tag{16.12}$$

$$a_i = \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,.,.}, \tag{16.13}$$

$$b_j = \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,.,.}, \tag{16.14}$$

$$c_k = \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{16.15}$$

$$d_{i,j} = \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,.,.}, \tag{16.16}$$

$$e_{i,k} = \overline{Y}_{i,.,k,.} - \overline{Y}_{i,.,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{16.17}$$

$$f_{j,k} = \overline{Y}_{.,j,k,.} - \overline{Y}_{.,j,.,.} - \overline{Y}_{.,.,k,.} + \overline{Y}_{.,.,.,.}, \tag{16.18}$$

$$g_{i,j,k} = \overline{Y}_{i,j,k,.} - \overline{Y}_{i,j,.,.} - \overline{Y}_{i,.,k,.} - \overline{Y}_{.,j,k,.} + \overline{Y}_{i,.,.,.} + \overline{Y}_{.,j,.,.} + \overline{Y}_{.,.,k,.} - \overline{Y}_{.,.,.,.}, \tag{16.19}$$

$$\epsilon_{i,j,k,l} = Y_{i,j,k,l} - \overline{Y}_{i,j,k,.}. \tag{16.20}$$

By relocating the overall mean term μ from the right side of Equation 16.3 to the left, we derive the following equation:

$$Y_{i,j,k,l} - \mu = a_i + b_j + c_k + d_{i,j} + e_{i,k} + f_{j,k} + g_{i,j,k} + \epsilon_{i,j,k,l}. \tag{16.21}$$

Squaring and summing across all responses for both sides of the equation, the left side represents the total variance of the responses, while the right side decomposes into the sum of squares (SS) for each effect. Notably, cross-product terms among distinct effects sum to zero due to orthogonal design properties. Thus, the total variance is additively partitioned into contributions from main effects, interactions, and random errors. Denoting SS as the sum of squares, we have the following equations:

$$SSY = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \sum_{l=1}^{r} Y_{i,j,k,l}^2, \tag{16.22}$$

$$SS0 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}\mu^2, \tag{16.23}$$

$$SST = SSY - SS0, \tag{16.24}$$

$$SSA = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}a_i^2, \tag{16.25}$$

$$SSB = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}b_j^2, \tag{16.26}$$

$$SSC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}c_k^2, \tag{16.27}$$

$$SSAB = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}d_{i,j}^2, \tag{16.28}$$

$$SSAC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}e_{i,k}^2, \tag{16.29}$$

$$SSBC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}f_{j,k}^2, \tag{16.30}$$

$$SSABC = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}g_{i,j,k}^2, \tag{16.31}$$

$$SSE = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3}\sum_{l=1}^{r}\epsilon_{i,j,k,l}^2. \tag{16.32}$$

So, we have:

$$SSY - SS0 = SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC + SSE. \tag{16.33}$$

To quantify the contribution of each effect to the response variance, we calculate the ratio of the specific effect's sum of squares (SS) to the total variance (SST). By dividing the sum of squares (SS) for each effect by its corresponding degrees of freedom (df), we obtain the mean squares (MS). The degrees of freedom (df) for each sum of squares are shown in Table 16.1.

We calculate the mean squares (MS) of effects as follows: So, we have:

$$MSA = \frac{SSA}{n_1 - 1}, \tag{16.34}$$

SS	*df*
SSY	$n_1 \times n_2 \times n_3 \times r$
SS0	1
SST	$n_1 \times n_2 \times n_3 \times r - 1$
SSA	$n_1 - 1$
SSB	$n_2 - 1$
SSC	$n_3 - 1$
SSAB	$(n_1 - 1) \times (n_2 - 1)$
SSAC	$(n_1 - 1) \times (n_3 - 1)$
SSBC	$(n_2 - 1) \times (n_3 - 1)$
SSABC	$(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)$
SSE	$n_1 \times n_2 \times n_3 \times (r - 1)$

Table 16.1: The degrees of freedom (df) for each sum of squares (SS).

$$MSB = \frac{SSB}{n_2 - 1}, \tag{16.35}$$

$$MSC = \frac{SSC}{n_3 - 1}, \tag{16.36}$$

$$MSAB = \frac{SSAB}{(n_1 - 1) \times (n_2 - 1)}, \tag{16.37}$$

$$MSAC = \frac{SSAC}{(n_1 - 1) \times (n_3 - 1)}, \tag{16.38}$$

$$MSBC = \frac{SSBC}{(n_2 - 1) \times (n_3 - 1)}, \tag{16.39}$$

$$MSABC = \frac{SSABC}{(n_1 - 1) \times (n_2 - 1) \times (n_3 - 1)}, \tag{16.40}$$

$$MSE = \frac{SSE}{n_1 \times n_2 \times n_3 \times (r - 1)}. \tag{16.41}$$

The F-statistic is calculated as the ratio of mean squares (MS) for any two effects among Equation 16.34 and Equation 16.41. By comparing this ratio to the critical F-value from the F-distribution with corresponding degrees of freedom (df), an F-test can assess the statistical significance of the two effects. For example, we calculate $F_c = \frac{MSA}{MSE}$. At a 0.05 significance level, we refer to the F-distribution with $df_1 = n_1 - 1$ numerator and $df_2 = n_1 \times n_2 \times n_3 \times (r - 1)$ denominator degrees of freedom to obtain the critical value F_v corresponding to 95% cumulative probability. If $F_c \geq F_v$, we conclude that the main effect of Factor A is statistically significant relative to random errors.

16.6 Examples: Infer the Effects of Apple Origins on Purchasing Prices

We then conduct a case study on the apple purchasing price, performing ANOVA and F-tests within the DoE framework.

Design of Experiment

Case study

The price of purchasing a box (5kg) of apples is related to the combination of its origin and size. There are two regions of origin: Qixian County in Henan Province (Origin 1) and Yantai City in Shandong Province (Origin 2), and three types of size: small (Size 1), medium (Size 2), and large (Size 3). The corresponding apple prices are shown in the table below.
We will analyze the effects of the origins and sizes of the apples on the purchasing price, and determine whether the origin or the size has a more statistically significant influence on the price. These factors and their corresponding levels in the example will be reused throughout all chapters of Part IV.

	Size 1	Size 2	Size 3
Origin 1	¥25	¥45	¥50
Origin 2	¥55	¥65	¥90

Table 16.2: Price of purchasing one box (5kg) of apples.

Calculation

The model equation of this example is:

$$Y_{i,j} = \mu + a_i + b_j + c_{i,j} + \epsilon_{i,j,k}.$$

Where μ is the overall mean effect, a_i is the effect of origin, b_j is the effect of size, $c_{i,j}$ is the effect of interaction between origin and size, $\epsilon_{i,j,k}$ is the effect of random errors. In this case, the price has no chance of fluctuations, so $\epsilon_{i,j,k} = 0$.
We calculate each mean as follows:

$$\overline{Y}_{.,.} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{3} Y_{i,j}}{2 \times 3},$$
$$\overline{Y}_{i,.} = \frac{\sum_{j=1}^{3} Y_{i,j}}{3},$$
$$\overline{Y}_{.,j} = \frac{\sum_{i=1}^{2} Y_{i,j}}{2}.$$

Then each effect in the model equation is:

$$\begin{aligned} \mu &= \overline{Y}_{.,.}, \\ a_i &= \overline{Y}_{i,.} - \overline{Y}_{.,.}, \\ b_j &= \overline{Y}_{.,j} - \overline{Y}_{.,.}, \\ c_{i,j} &= Y_{i,j} - \overline{Y}_{i,.} - \overline{Y}_{.,j} + \overline{Y}_{.,.}. \end{aligned}$$

We can then calculate the sum of squares of each effect as follows:

$$\begin{aligned} SSY &= \sum_{i=1}^{2}\sum_{j=1}^{3} Y_{i,j}^2 = 20,500, \\ SS0 &= \sum_{i=1}^{2}\sum_{j=1}^{3} \mu^2 = 18,150, \\ SST &= SSY - SS0 = 2,350, \\ SSA &= \sum_{i=1}^{2}\sum_{j=1}^{3} a_i^2 = 1,350, \\ SSB &= \sum_{i=1}^{2}\sum_{j=1}^{3} b_j^2 = 900, \\ SSAB &= \sum_{i=1}^{2}\sum_{j=1}^{3} c_{i,j}^2 = 100. \end{aligned}$$

Factor	Contribution
Origin	57.45%
Size	38.30%
Origin-Size	4.25%

Table 16.3: ANOVA of the price of a box (5kg) of apples.

Conclusion

For this example, the degrees of freedom for each factor are shown in the following Table.

Factor	df
Origin	1
Size	2
Origin-Size	2

Table 16.4: Degrees of freedom of each factor .

Then we can calculate the mean of the square of each effect as follows:

$$MSA = \frac{SSA}{df_a} = \frac{1,350}{1} = 1,350,$$
$$MSB = \frac{SSB}{df_b} = \frac{900}{2} = 450,$$
$$MSAB = \frac{SSAB}{df_{ab}} = \frac{100}{2} = 50.$$

We compare the origin to the size for the price of purchasing a box (5kg) of apples with a significance of 95%. The computed F-value f-compute is:

$$f\text{-compute} = \frac{MSA}{MSB} = \frac{1,350}{450} = 3.$$

The critical value in the F-distribution with 1 numerator and 2 denominator degrees of freedom at 95% significance is 18.51.

$$f\text{-table} = 18.51,$$
$$f\text{-compute} < f\text{-table}.$$

Therefore, at the 95% significance, we cannot conclude that origin is more statistically significant than size for the price of a box (5kg) of apples.

16.7 Limitations

DoE is a valuable tool for analyzing the effect relationships between responses and factor effects, enabling the calculation of the variance contribution of different factors to the response as well as their statistical significance. However, DoE also has limitations that researchers should be mindful of when applying it.

Firstly, DoE requires exhaustive experimentation across all experimental condition groups composed of specific level values for each investigated factor. Without responses from any single experimental condition group, ANOVA and F-tests cannot be conducted.

After selecting factors and their level values, a grid can be plotted where factors represent different dimensions, and experimental condition groups correspond to points on the grid. During experimentation, responses measured for each experimental condition group are then filled into their respective positions on the grid. Any missing response data at any point on the grid precludes ANOVA and F-tests.

Secondly, as the number of factors selected in DoE increases, the costs associated with conducting experiments and analyzing data grow exponentially. Moreover, ANOVA and F-tests, which require calculating the sum of squares (SSE) of responses across different experimental condition groups, incur higher computational costs compared to methods that simply calculate means. Consequently, the experimental costs of DoE remain a critical factor to be balanced.

Finally, conducting ANOVA and F-tests in DoE requires testing that the corresponding statistical assumptions are met, which necessitates validating whether these assumptions hold for the response variables.

16.8 Summary

The DoE employs a model equation to characterize the effects between factors and responses. Its model equation quantifies: 1) the overall mean of all responses, 2) main effects for individual factors, 3) factors' interaction effects 4) random errors due to chance fluctuations. By grouping experimental responses according to factor levels and calculating corresponding group means, component values for each model equation term can be estimated. The SS for all modeled effects (including errors) partitions total response variance, enabling ANOVA and an F-test to be conducted for statistical analysis.

Chapter 17

Randomized Controlled Trials

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of Randomized Controlled Trials (in short, RCTs).

17.1 Basic Concepts

The definitions of the basic concepts are cited from the early work of RCTs [191, 141, 59, 60].

An Intervention or Treatment is the precisely defined procedure whose effects or safety are under investigation. This can be a pharmacological agent (drug), a medical device, a surgical technique, a behavioral therapy, an educational program, or any other maneuver administered to participants to elicit a measurable response.

A Participant or Unit is an individual who meets the predefined eligibility criteria for a study and is formally enrolled to undergo the investigative procedures, data collection, and follow-up as outlined in the trial protocol.

Outcomes are quantitative metrics obtained from the participants through measurements, distinct from categorical variables.

A Treatment Group is the cohort of participants who are randomly allocated to receive the treatment, which is the primary focus of the investigation.

A Control Group is the cohort of participants who are randomly allocated to receive a comparator against which the treatment is evaluated. This comparator can be a placebo (an inert substance), a standard-of-care treatment, a different treatment, or no treatment, depending on the research question and ethical considerations.

Randomization means every participant has an equal chance of being assigned to either a Treatment Group or Control Group.

An Investigator is a qualified individual responsible for the conduct of the RCTs at a trial site. The Principal Investigator (PI) bears the overall responsibility for the ethical and protocol-directed execution of the study, including management of the investigative team, and protection of participant rights and data integrity.

[1] Mr. Chenxi Wang is the primary contributor.

The *Controlled Comparison* is a fundamental methodological principle in RCTs whereby the outcomes of the Treatment Group are systematically evaluated against those of the Control Group. The primary objective is to isolate the causal effect of the treatment by holding constant, through the research design, the influence of all other extraneous variables.

A Blinding is a methodological procedure in RCTs whereby one or more parties involved in the research are kept unaware of the treatment assignment (i.e., whether a participant is in the Treatment Group or the Control Group). The primary purpose is to prevent the introduction of conscious or unconscious bias that could influence the perceived outcomes, behaviors, or interpretations of the results.

The *Average Treatment Effect, in short ATE,* is the expected effect difference in outcomes between the Treatment Group and the Control Group. For any participant i, there are two potential outcomes: the observed outcome $Y_i(1)$ if assigned to the Treatment Group and the observed outcome $Y_i(0)$ if assigned to the Control Group. However, both potential outcomes for the same participant i cannot be observed simultaneously. RCTs utilize randomization to calculate an unbiased estimator of the average treatment effect.

The *External Validity* refers to the extent to which the results of RCTs can be generalized and applied beyond the research experimental setting (to real-world setting), and this extent can be evaluated through the correlation between RCT results and those obtained in real-world setting.

The *Internal Validity* refers to the degree to which RCTs can reliably demonstrate the results, meaning that the calculated ATE is indeed attributable to the difference between the treatment and control procedure, rather than to other confounding factors. By employing randomization, controlled comparison between Treatment Group and Control Group, and blinding, RCTs ensure the internal validity.

17.2 Problem Statement

RCTs are a systematic methodology for inferring the relationship between interventions and outcomes [191, 141].

Given a population of participants and a proposed treatment as inputs, each participant i possesses two potential outcomes, $Y_i(1)$ when receiving the treatment and $Y_i(0)$ when receiving the control. The challenges lie in that only one of these outcomes can ever be observed for the same participant i. Please note that in several cases, we could observe both outcomes, as we have discussed in Section 14.

The RCTs problem could be formulated as "how to use a random mechanism (randomization) to assign participants to either the Treatment Group or Control Group, providing an unbiased estimate of the Average Treatment Effect (ATE)." The inputs for RCTs are participants, the Treatment Group, the Control Group, and the observed outcome in each participant. Under the constraint of randomly assigning participants into the Treatment Group and Control Group, RCTs output the ATE.

17.3 Basic Assumptions

Assumption 17.1 (Stable Unit Treatment Value Assumption (SUTVA) [43, 158])**.** *The potential outcomes for any unit or participant, e.g., an individual patient, are unaffected by the particular treatment assignment of other units or participants. This assumption comprises two key components:*

1. *No Interference: The treatment assignment of one unit does not influence the outcome of any other unit. This ensures the independence of units, a prerequisite for standard statistical inference.*

2. *No Hidden Variations of Treatment: Each treatment regime is identical for every unit assigned to it. There are no multiple, unspecified versions of the treatment or control that could differentially affect the outcomes.*

SUTVA is the most fundamental assumption underpinning RCTs, as it allows us to define a unique potential outcome for each unit under each treatment state.

Assumption 17.2 (Ignorability)**.** *This assumption is also known as unconfoundedness or exchangeability. It states that the assignment of treatment is conditionally independent of the potential outcomes, given the observed covariates. Formally,*

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X, \tag{17.1}$$

where $Y(1)$ and $Y(0)$ are the potential outcomes under the Treatment Group and the Control Group, respectively, T is the treatment assignment indicator, and X represents a vector of pre-treatment covariates. The symbol "$\mid$" is a mathematical notation denoting "under the condition of," and "$A \mid B$" signifies "A under the condition of B." The symbol "$\perp\!\!\!\perp$" represents "mutual independence," with "$A \perp\!\!\!\perp B$" indicating that "A and B are mutually independent."

Randomization is the physical design feature that ensures this assumption is met. It renders the Treatment Group and the Control Group statistically equivalent, or exchangeable, in expectation, not only with respect to measured covariates X but also with respect to all unmeasured factors. Consequently, any systematic difference in observed outcomes can be causally attributed to the treatment.

Assumption 17.3 (Consistency)**.** *The observed outcome for a participant who received a specific treatment level is precisely the potential outcome for that participant under that same treatment level. Formally, if $T_i = t$, then $Y_i^{obs} = Y_i(t)$.*

This assumption implies that the treatment is well-defined and that there is no ambiguity in its application or measurement across all participants. It links the conceptual potential outcomes to the actually observed data.

17.4 Basic Principles

The primary motivation for employing RCTs lies in the need to eliminate confounding bias by ensuring that the Treatment Group and Control Group are statistically equivalent at baseline. This is achieved through the random assignment of participants to experimental conditions, thereby ensuring that both observed and unobserved covariates are, on average, balanced across groups.

RCTs typically involve the comparison of one or more treatment interventions against a control condition, often incorporating blinding and placebo controls to minimize expectancy and measurement biases. The methodology is characterized by a clearly defined protocol, prespecified outcomes, and rigorous adherence to statistical principles such as intention-to-treat analysis. Widely applied in clinical medicine, social sciences, and public policy evaluation, RCTs enable researchers to make high-confidence claims about the effect of interventions under controlled and replicable conditions[191].

Randomization is the heart of the RCTs. It ensures that every participant has an equal chance of being assigned to either the Treatment Group or the Control Group. This is crucial because it ensures the groups are, on average, similar in every way at the start of the study, not just in age or gender, but also in countless other factors we can't even measure (like genetics, diet, or natural resilience). RCTs control for other factors (known and unknown) beyond the Treatment/Control Group through randomization, reducing effects to a matter of chance fluctuations.

Because the groups are comparable, any significant difference in the outcome at the end of the study can be more confidently attributed to the treatment itself, rather than to pre-existing differences between the groups.

17.5 Methodology

17.5.1 The Average Treatment Effect Formulation

We can formally define the controlled comparison made in RCTs. The goal is to estimate the Average Treatment Effect (ATE), which is the average causal effect of the treatment across the entire study population.

We begin by defining the key steps in RCTs:

- Let $i = 1, 2, ..., N$ represent each of the N total participants in the trials.
- The treatment assignment for Participant i is denoted by T_i:

$$T_i = \begin{cases} 1, & \text{if Participant } i \text{ is assigned to the Treatment Group,} \\ 0, & \text{if Participant } i \text{ is assigned to the Control Group.} \end{cases} \tag{17.2}$$

- If $T_i = t$, the observed outcome for Participant i (e.g., the decrease in apple purchasing price) is denoted by $Y_i(t)$.

Based on the above discussion, the observed outcome for Participant i depends on the treatment assignment:

$$Y_i^{obs} = Y_i(T_i) = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0). \tag{17.3}$$

The fundamental advantage of randomization is that it allows for a simple and powerful comparison. The Average Treatment Effect (ATE) is estimated by calculating the difference in the average outcome between the two groups:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)], \tag{17.4}$$

where:

- $E[Y_i(1)]$ is the expected value (arithmetic average) of the outcomes for all participants in the Treatment Group.
- $E[Y_i(0)]$ is the expected value of the outcomes for all participants in the Control Group.

In practice, we calculate the sample averages to estimate the ATE:

$$\widehat{ATE} = \frac{1}{N_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{N_0} \sum_{i:T_i=0} Y_i(0). \tag{17.5}$$

Where:

- N_1 is the number of participants in the Treatment Group.
- N_0 is the number of participants in the Control Group.
- $N = N_0 + N_1$

17.6 Example: Calculate the Effects of Apple Origins on Purchasing Prices

We then conduct a case study of performing RCTs to investigate the apple origin trial on the purchasing price.

Randomized Controlled Trails

Case study

- Presume that 10,000 ($N = 10,000$) apple seeds in your study.
- You use a computer to randomly assign 5,000 ($N_1 = 5,000$) apple seeds for planting in Qixian County, Henan Province (*Treatment Group*).
- The other 5,000 ($N_0 = 5,000$) seeds were planted in Yantai City, Shandong Province (*Control Group*).

After 10 years, you measure the average purchasing price per kilogram for the apples grown in both groups.
Suppose the obtained average purchasing price per kilogram is shown in the following Table.

Group	T_i	$E[Y_i(T_i)]$
Treatment Group	1	¥8/kg
Control Group	0	¥14/kg

Table 17.1: The average purchasing price per-kilogram in two trials.

Calculation

Applying the Formula to the Example:

- The average purchasing price per-kilogram in the Treatment Group ($E[Y_i(1)$) was ¥8/kg.
- The average purchasing price per-kilogram in the Control Group ($E[Y_i(0)]$) was ¥14/kg.
- Using Equation 17.5: $\widehat{ATE} = 8 - 14 = -6$.

Conclusion

This suggests apples grown in Qixian County, Henan Province had an average purchasing price of ¥6/kg *lower* than that of apples grown in Yantai City, Shandong Province. Because of randomization, we can be confident that this difference is likely due to the origin itself and not other confounding factors.

17.7 Limitations

RCTs are widely regarded as the gold standard for establishing causal inference, yet they are subject to several important limitations that researchers must acknowledge.

Firstly, the high costs and complex logistics associated with properly conducting an

RCT often pose significant barriers. Implementing randomization, ensuring adherence to protocols, and maintaining blinding over a sufficient follow-up period require substantial financial resources, time, and operational effort, which can be prohibitive for many research questions.

Secondly, ethical and practical constraints frequently limit the applicability of RCTs. It is ethically impermissible to randomize participants to interventions known or strongly believed to be harmful. Furthermore, for research on long-term outcomes or rare diseases, it may be practically infeasible to recruit and retain an adequate sample size over the necessary timeframe.

Thirdly, the issue of generalizability (external validity) is a critical concern. The highly controlled conditions and specific participant population of an RCT may not be representative of real-world settings or broader patient groups. Consequently, findings from an RCT, while internally valid, may not translate directly to routine clinical practice.

Finally, RCTs are vulnerable to specific methodological biases if not meticulously designed and executed. These include attrition bias, if dropout rates (the proportion of participants who leave the study) differ systematically between groups, and a failure of blinding, which can introduce performance and detection bias. The analysis must also adhere to the intention-to-treat principle (analyzing all participants in their originally assigned groups regardless of what treatment they actually received) to preserve the integrity of the randomization, which can complicate the interpretation of the actual treatment effect received.

17.8 Summary

The RCTs are considered the gold standard for a simple reason: the design provides the clearest possible answer to the question “What would have happened otherwise?” The Control Group shows what happens in the absence of the intervention, and randomization ensures this comparison is fair. While other study designs can find associations, a well-conducted RCTs provides the strongest evidence for a causal relationship. This makes it an indispensable tool for advancing reliable knowledge in science and medicine. However, it is important to note that RCTs require a sufficient sample size due to randomization to reliably estimate causal effects.

Chapter 18

Quasi-experiments

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the quasi-experiments.

18.1 Basic Concepts

We will reuse most concepts defined in Chapter 17, and only introduce those concepts in quasi-experiments different from those of RCT.

A Treatment/ Controlled Group denotes the intervention or manipulation that is applied to the participants or groups. The explicit manipulation of this variable is the characteristic that links quasi-experimental research to real-world experiments. There are slight differences between the concepts of Treatment/Controlled in quasi-experiments from those of RCTs. In quasi-experiments, *a control Group is frequently nonequivalent because participants were not assigned through randomization.* According to Campbell et al. [29], "the concept of the application of an experimental mode of analysis and interpretation to bodies of data not meeting the full requirements of experimental control because experimental units are not assigned at random to at least two 'treatment' conditions." [2]

Cause refers to the differences between the Treatment Group and the Control Group in quasi-experiments on pre-existing groups, while *effect* refers to the differences in outcomes between the Treatment Group and the Control Group in the context of quasi-experiments.

Internal validity refers to the influence in a certain specific experimental instance.

External validity refers to how the influence can be generalized to populations, settings, treatments, and measurements [29].

Natural Group refers to research units that are pre-existing organizational structures in the real world [119]. Researchers are constrained to assign the treatment to the

[1] Mr Hongxiao Li is the primary contributor.

[2] According to the original author's intention, the 'two treatment conditions' should be the treatment group and the control group.

entire unit because they cannot subdivide these natural groups into randomly equivalent subgroups for either treatment or control. The use of natural grouping is a physical constraint that necessitates a quasi-experimental approach due to selection biases from the real world.

Confounding Variable is not strictly defined in the existing quasi-experiment studies. According to the description of Thyer et al. [203], confounding variables are referred to as other variables than the control and treatment, which are external and might make changes to the outcome if left uncontrolled on the dependent variable, thereby undermining the ability to claim a causal relationship between the treatment and the outcome. As a limitation, a quasi-experiment cannot completely infer the causal effect.

18.2 Problem Statement

Quasi-experimental designs [150] are a research design used to estimate causal relationships when random assignment of participants to experimental groups is not feasible or ethical. Unlike RCTs, it lacks full control over variables but still compares groups exposed to different conditions.

A quasi-experiment problem can be stated as "when a random mechanism (randomization) is not available, how to estimate the Average Treatment Effect (ATE)." Especially, quasi experiments focus on the outcome in a time period.

Same as RCTs, the inputs for quasi-experiments are participants, the treatment group, the control group, and the observed outcome in each participant. The outputs are the approximated effects of the Treatment Group rather than its causal effects.

18.3 Basic Assumptions

The basic assumptions of quasi-experiments are mostly the same as those of RCT, with the two major exceptions as follows:

Assumption 18.1 (Non-independence [29])**.** *Most quasi-experiments are conducted in real-world scenarios, where the correlation between individuals is stronger. This makes the assumption of no interference harder to guarantee.*

Meanwhile, the implementation of treatments may vary across scenarios (e.g., the same policy implemented in different hospitals), leading to a higher risk of violating the requirement of no hidden treatment variation.

Assumption 18.2 (Approximate unconfoundedness [29])**.** *There is no random assignment process. The formation of Treatment and Control Groups is usually related to individual characteristics, leading to the existence of confounding variables. So the unconfoundedness assumption is not fully satisfied.*

18.4 Basic Principles

Quasi-experiments provide a structured framework for approximating causal effects where randomization is infeasible or unethical. The principles of quasi-experiments explicitly distinguish between causal and associational relationships by carefully identifying sources of variation in a way that mimics randomization. These frameworks incorporate prior domain knowledge to design and justify the selection of comparison groups, ensuring that they approximate counterfactual scenarios. Quasi-experiments estimate the effect of intervention by leveraging natural or real-world variations to estimate the effects of interventions.

18.5 Methodology

Quasi-experimental designs [150] evolve from the simple before-and-after design through the following five key strategies, each targeting specific threats to internal validity.

1. *Add a non-randomized control group* [150]. This method is like RCTs. However, the partition is sometimes impossible. Therefore, a quasi-experiment setting is used.

In this and the following parts, we use Notation O to represent a measurement, and Notation X to represent an intervention. If there are two lines, they represent two groups. Corresponding to Notation X in a group, a blank in another group represents a non-intervention. This method can be represented as follows:

O X O
O O

Example [150]: To lower reindeer herders' high injury rates, three geographically divided groups were set: Group A got preventive measure letters, Group B got info from trained health staff, and the control group got nothing.

For a pre-period (1985) and a post-period (1987), the accident rates dropped for all: A:18.7→15.1; B:21→14.9; control:19.2→14.6. According to the standard deviation threshold requirement, no significant group difference emerged, so the interventions were deemed ineffective.

2. *Take more measurements* [150]. Instead of a single before-and-after measurement, multiple baseline measurements establish a pre-intervention, and multiple post-intervention measurements establish a post-intervention. Evaluation on several more measurements is more reliable.

This method can be represented as follows:

O O O X O O O

shows an example for a single before-and-after measurement and

O O O X O O O
O O O O O O

shows an example for multiple baseline measurements.

Example [150]: A food factory did safety training for two departments. It first took 3-4 weeks for safety behavior baseline checks (via checklist) for both departments. Then, after the post-intervention, including education, safety lists, and feedback, checks contin-

ued. The results showed safety behaviors changed after intervention in each department, strengthening the links between interventions and behaviors.

3. *Stagger the introduction of the intervention* [150]. The term "stagger" means all partitioned groups eventually receive the intervention, but at different times, with each group serving as a comparison for others. This minimizes chronological effects: coincidental events aligning with one intervention are plausible, but multiple such coincidences are unlikely.

For instance, the interventions of salary improvement for different groups are applied at different times. Therefore, the effects' discrepancy of salary improvement is less interfered with by occasional incidents, e.g., natural disasters or economic fluctuations. This method can be represented as follows:

O O O X O O O O O O
O O O O O O X O O O

Example [150]: The same food factory launched safety training first in the wrapping department, then later in the make-up department. Each had baseline checks before training. A staggered launch lets departments act as comparisons, reducing the chance of extraneous events causing results.

4. *Reverse the intervention* [150]. Compare the effects of the intervention and that after reversing the intervention. If outcomes revert to baseline levels after reversal, the intervention's causality is confirmed. This method eliminates chronological and occasional threats and errors involved in the testing procedure, but it is only suitable when the intervention is reversible. With the symbol $-X$ standing for the reversed intervention, this method can be represented as follows:

O O O X O O O -X O O O

Example [150]: A company plans a participatory ergonomics program with task modifications, new equipment, and worker education. To measure its impact, the coordinator will compare symptoms and injuries before and after, while tracking equipment purchases and self-reports on tasks and stressors to rule out other factors.

5. *Measure multiple outcomes* [150]. There are two approaches. The first one is to add intervening outcome measures. Track implementation and short or intermediate-term effects to distinguish between ineffective interventions and flawed implementation. For example, a "train-the-trainer" program's lack of injury reduction was attributed to poor implementation rather than ineffective techniques. Second, add related but untargeted outcomes. Measure outcomes similar to the main target but unaffected by the intervention. This method can be represented as follows, where O_1/O_2 stands for two measured quantities:

$$O_1/O_2 \text{ X } O_1/O_2$$

Example 1 [150]: Add intervening measures: A company's ergonomics program tracked ultimate outcomes (symptoms/injuries) and added checks like equipment purchase records and work task reports. This helped tell if failure was from a bad program or poor implementation.

Example 2 [150]: Add untargeted measures: Oil platforms tracked tong-related injuries (target of new equipment) and non-tong injuries (untargeted). No change in

non-tong injuries meant that the tong injury changes were likely from the equipment. A grocery ergonomic intervention found targeted (neck/back) discomfort dropped, but untargeted (arm/wrist) didn't, reducing validity threats.

Besides the above, analytical methods such as matching, propensity score, and difference-in-differences are required to control for observable confounders as much as possible, but unobservable confounders cannot be addressed.

18.6 Example: The Impact of Origin on Apple Purchasing Price

Quasi experiments

Case study

This example uses the first strategy. To investigate how the apple origin influences market prices, experiments ought to be conducted across two origins (designated as the control group and treatment group). Nevertheless, given the insufficient sample size of the selected origins (due to policy or practical constraints), supplementary sample data were collected both from the origins themselves and their adjacent areas. The design leverages a natural characteristic: for the origins with a small sample size, geographically adjacent areas —often sharing identical climate, soil, and farming practices—can serve as their reference. The quasi-experiment design approximately ensures that the only systematic difference between the two groups is "origin label (X)", while confounding factors (A, B, C) remain balanced, where A stands for sweetness (12, 14, 16 Brix), B stands for apple size (small, medium, large), and C stands for shelf-life (3 days, a week, two weeks).

1. *Treatment Group*
   - Scope: Apple growers located within 5 km of the Origin 1.
   - Label: Apples are marketed with the target origin certification ($X = 1$).

2. *Control Group*
   - Scope: Apple growers located within 5 km of the Origin 2.
   - Label: Apples are marketed with the target origin certification ($X = 2$).

Table 18.6 is the price per kilogram of apples from two origins.

Group	T_i	$E[Y_i(T_i)]$
Treatment Group	1	¥8/kg
Control Group	0	¥14/kg

Table 18.1: The average apple per-kilogram purchasing price in two trials.

Calculation

Before analyzing the price effect, it is necessary to verify that A, B, C are balanced between the two groups, to rule out the impact of non-origin factors on price.

1. *Data Collection*
   - For each sampled apple orchard, collect 50 representative apples.
   - Measure indicators: Sweetness (A, via Brix refractometer), size (B, via single fruit weight in grams), shelf life (C, via days of normal storage without decay).

2. *Balance Test Methods*
   - Descriptive statistics: Calculate the mean, standard deviation, and median of A, B, C for both groups.
   - Statistical significance test: Use an independent samples t-test to verify that the mean difference of A, B, C between the two groups is not statistically significant (standard threshold: p-value > 0.05).
   - Distribution visualization: Draw kernel density plots for A, B, C; overlapping curves indicate balanced distributions.

After confirming the balance of confounding factors, the effect of origin (X) on purchase price (Y) is calculated through direct comparison:

1. *Price Data Collection*
   - Track the same batch of merchants to collect on-site purchase prices of apples from both groups.
   - Record Y as the unit price of apples meeting the same basic quality standards.

2. *Effect Calculation*
   1. Calculate the average purchase price of the treatment group ($\bar{Y}_1$) and the control group ($\bar{Y}_0$).
   2. Compute the effect: Effect $= \bar{Y}_1 - \bar{Y}_0 = 8 - 14 = -6$.

   Explanation: The value represents the additional price brought by the "target origin label", as confounding factors (A, B, C) have been balanced by geographic matching.

Conclusion

The result reports apples grown in Origin 1 had an average purchasing price of ¥6/kg *lower* than that of apples grown in Origin 2. Therefore, it is confident that this difference is likely due to the origin itself and not other confounding factors.

18.7 Limitations

The primary limitation of quasi-experiments is the lack of internal validity, let alone external validity. The lack of internal validity comes from the following multiple factors.

1. *Lack of randomization*: In quasi-experiments, objects are not randomly assigned to Experimental and Control Groups. Grouping is based on pre-existing conditions as a weak substitute. This lack of randomization leads to selection bias, where systematic differences between the groups may exist independently of the Treatment and Control.

2. *Effect of confoundings*: Without full control over the experimental environment, other confoundings may simultaneously interfere with both the independent and dependent variables, masking or exaggerating the true effect of the independent variable.

3. *History effects*: As quasi-experiments are often deployed in real-world scenarios, some external events that occur during the study can influence the dependent variable, regardless of the experimental treatment. These events can confound the results. For example, a policy intervention during a specific period might be interfered with by simultaneous economic changes.

18.8 Summary

Quasi-experiment is a research design that approximates the methodological rigor of a true experiment. Quasi-experiments enable researchers to approximate the effects of the Treatment group. This distinguishes them from purely observational studies, as quasi-experiments often involve deliberate manipulation of an independent variable or exploitation of natural interventions. Although a quasi-experiment is useful in social science and some other fields, its limitations of internal validity and generalization failure are inherent. The approximated effects are not a rigorous causal effect.

Chapter 19

Structural Causal Model

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the structural causal model (in short, SCM).

19.1 Basic Concepts

Endogenous variables are the observed variables of interest.

Exogenous variables originate outside the model and are not influenced by any other variables within the system..

A directed acyclic graph (DAG) is a representation used in graph theory where nodes represent variables and directed edges represent causal directionality [202, 12]. The graph must be acyclic, meaning no path of arrows can lead back to a node already included in that path. The absence of an arrow between two variables constitutes an empirical assumption.

An intervention is mathematically formalized using the $\mathbf{do}(X = x)$ operator, which refers to a physical action setting the variable X to a fixed value x in its value range. Pearl et al. first formalized this concept in the definition of Causal Bayesian Network [139]. Deleting the structural equation for X and replacing it with the fixed value equation $X = x$ in the original SCM, noted as M, results in a modified sub-model M_x.

Causal effect is defined as the probability distribution $\mathbb{P}(Y = y|\mathbf{do}(X = x))$, which is the distribution of Y in sub-model M_x. This concept is implicitly defined by Pearl et al. [139] in the original text as: "A causal structure or relationship or mechanism serves as a blueprint for forming a 'causal model' – a precise specification of how each variable is influenced by its parents in the DAG."

A direct cause is implied by a direct causal relationship $X \to Y$ that exists if X is a parent of Y in the causal graph and appears as an argument in the structural equation for Y. This relationship quantifies the causal effect of X on Y, asserting that a change in X results in a corresponding change in Y regardless of the values taken by other variables in the model. The variable Y is called the effect of X in this context [139, 141].

[1] The primary contributor is Mr. Hongxiao Li.

A potential cause X *of* Y is defined graphically, where X is an ancestor node of Y, meaning a directed path exists between them. Operationally, X is considered a cause of Y if the intervention $\mathbf{do}(X = x)$ alters the probability distribution of Y, $\mathbb{P}(Y|\mathbf{do}(X = x))$, thus allowing for the inference of relationships that remain invariant under external change.

The following is the original formalization of Pearl et al. [139]:
"A variable X has a potential causal influence on another variable Y (that is inferable from P) if the following conditions hold:

1. X and Y are dependent in every context.
2. There exists a variable Z and a context S such that:
   (i) X and Z are independent given S (i.e., $X \perp\!\!\!\perp Z \mid S$). The notation $\perp\!\!\!\perp$ means independent.
   (ii) Z and Y are dependent given S (i.e., $Z \not\perp\!\!\!\perp Y \mid S$)." The notation $\not\perp\!\!\!\perp$ means dependent.

Counterfactual refers to the value Y "would have taken" under a hypothetical condition $X = x$, are defined using the intervention sub-model M_x [139] by Pearl et al. The counterfactual $Y_x(u)$ with background factors $U = u$ is defined as the value of Y computed in the sub-model M_x: $Y_x(u) \triangleq Y_{M_x}(u)$. This formulation means that the probability of a counterfactual $\mathbb{P}(Y_x = y)$ is the probability assigned to the set of variables U that satisfy $Y_{M_x}(u) = y$. The symbol $\triangleq$ is used for definition.

19.2 Problem Statement

SCM [89, 141] is a foundational statistical and conceptual framework that uses causal graphs and structural equations to represent and infer causal relationships between variables. This approach is widely used in fields such as epidemiology, economics, computer science, and social sciences for tasks like causal discovery, mediation analysis, and policy evaluation.

The SCM problem could be stated as "how to model the causal mechanisms underlying a system in what formal language to enable the estimation of causal effects that go beyond mere correlational patterns observed in data."

The inputs to this problem consist of a DAG encoding the causal structure, and a distribution over exogenous noise variables. The output is the estimation of causal effects among those inputs. The constraints require that the graph must be acyclic, each variable has a unique structural equation, and exogenous noise terms satisfy specified independence assumptions.

19.3 Basic Assumptions

The structural causal model method relies on the following three assumptions.

Assumption 19.1 (Causal DAG [139])**.** *It is assumed that the causal relationships between variables can be represented by a DAG, and the graph is acyclic (i.e., no variable can influence itself through a causal path).*

SCMs are represented using a directed acyclic graph (DAG), where:

- *Nodes represent variables.*
- *Directed edges represent causal relationships.*

Assumption 19.2 (Structural Equations [139])**.** *It is assumed that each variable X_i is determined as a function of its direct causes (parents in the DAG) and an independent noise term U_i:*

$$X_i = f_i(\mathbf{Pa}(X_i), U_i), \tag{19.1}$$

where:

- $\mathbf{Pa}(X_i)$ *represents the direct causes (parents) of X_i in the DAG.*
- f_i *is a deterministic function describing how X_i depends on its parents.*
- U_i *is an independent noise term capturing exogenous influences.*

Assumption 19.3 (Causal Markov Condition [78, 139])**.** *It is assumed that each variable X_i is conditionally independent of its non-descendants (any other variables) given its parents in a causal DAG:*

$$X_i \perp\!\!\!\perp \textit{Non-Descendants of } X_i \mid \mathbf{Pa}(X_i). \tag{19.2}$$

This assumption allows the global probability distribution to be factorized based on the local causal relationships.

19.4 Basic Principles

SCM provides a formal language for encoding causal assumptions, deriving causal inferences, and quantifying the effects of interventions. SCM moves beyond purely statistical associations by explicitly incorporating causal knowledge through a structured system of equations and a corresponding graphical representation.

The formal language explicitly distinguishes between causal and associational relationships, encodes prior domain knowledge into the structure of causal graphs where edges represent direct causal effects, defines and computes counterfactuals through interventions that simulate external manipulations of variables, and answers "what-if" questions. These frameworks typically involve a set of structural equations that quantify how each variable is determined by its direct causes and unobserved noise terms, which capture unmeasured factors and random variation, while also clarifying the distinction between confounding variables that affect both cause and effect and mediators/moderators that transmit or modify causal effects within a coherent causal hierarchy.

19.5 Methodoology

In this section, big letters stand for spaces and small letters stand for the element in the spaces.

19.5.1 Abilities of SCM

SCM is formalized using a triple (V, W, F), where:

- V is a set of endogenous variables.
- W is a set of exogenous variables, which are not caused by any variable in V.
- F is a set of functions $f_1, f_2, \cdots, f_n$ that assign each variable $v_i \in V$ a value based on the values of its causal parents and an exogenous variable W_i, i.e., $v_i = f_i(pa_i, w_i)$.

It provides the following abilities.

- *Explicit Causal Assumptions:* The graphical model forces researchers to make their causal assumptions transparent and testable. The implications of these assumptions (e.g., conditional independencies) can then be checked against the data.
- *Handling of Confounding:* SCM provides a clear framework for defining, identifying, and adjusting for confounding variables, which are common sources of bias in observational studies.
- *Unification of Concepts:* SCM unifies concepts from statistics, graph theory, and potential outcomes into a single coherent framework, facilitating a deeper understanding of causality.
- *Counterfactual Reasoning:* It provides a formal semantics for answering counterfactual questions, which are central to explanation, attribution, and legal reasoning.

19.5.2 Fundamental Methodology

Basic Equations: The core of an SCM is its structural equations. For each endogenous variable V_i, we have:

$$V_i := f_i(PA_i, W_i), \tag{19.3}$$

where PA_i are the direct parents of v_i in the graph and W_i is an exogenous variable. The assignment operator $:=$ signifies a causal assignment, not just a mathematical equality.

The do-operator and Intervention: The effect of an intervention that sets a variable X to a value x is represented by the *do*-operator, $\mathbf{do}(X = x)$. This operation modifies the SCM by replacing the structural equation for X with the equation $X = x$. The resulting probability distribution, denoted as $\mathbb{P}(Y|\mathbf{do}(X = x))$ or $\mathbb{P}_x(y)$, is the *interventional distribution*.

Identification Condition: A central question in SCM is whether a causal effect, $\mathbb{P}(Y|\mathbf{do}(X = x))$, can be uniquely determined from the observed data distribution and the causal graph G. This is the problem of *identification*. A key identifiability result is the *back-door criterion*: if a set of variables Z satisfies the back-door criterion relative to (X, Y) (i.e., Z blocks all back-door paths from X to Y and contains no descendants of X), then the causal effect is identified by the adjustment formula:

$$\mathbb{P}(Y|\mathbf{do}(X = x)) = \sum_z \mathbb{P}(Y|X = x, Z = z)\mathbb{P}(Z = z). \tag{19.4}$$

The formula represents the causal effect of an intervention $\mathbf{do}(X = x)$ on Y. It can be interpreted as a weighted sum over all possible values of the mediator variable Z. The formula is composed of two parts:
1. $\mathbb{P}(Y|X = x, Z = z)$: The conditional probability of Y given $X = x$ and $Z = z$.
2. $\mathbb{P}(Z = z)$: The natural distribution of Z without any intervention.

This formula shows that the causal effect $\mathbb{P}(Y|\mathbf{do}(X = x))$ can be computed by adjusting for the mediator Z. It is commonly used in causal inference to estimate intervention effects.

Another powerful identifiability tool is the *front-door criterion*, which can be used even in the presence of unmeasured confounders.

If there is no confounding variables that have paths to both X and Y, we call paths from X to Y textitfront-door paths. In this case, adjustment on variable Z is not feasible because it changes the distribution of X and Y. In this case, the causal effect is computed with the following formula according to the total probability theorem:

$$\mathbb{P}(Y|\mathbf{do}(X = x)) = \mathbb{P}(Y|X = x) = \sum_z \mathbb{P}(Y|X = x, Z = z). \tag{19.5}$$

19.5.3 Procedures

Evaluations using the SCM method follow the following six steps. The fifth step is essential.

1. **Define the Problem**: Define the research problem, factors, and causal assumptions.
2. **Construct a Causal Graph**: Use a DAG to represent causal relationships among variables, marking confounders.
3. **Determine Identifiability**: Check if causal effects can be estimated.
4. **Data Processing**: Collect data and handle missing values, outliers, etc.
5. **Causal Inference**: Compute causal effect using the structure formula, for example 19.4 or 19.5.
6. **Validation and Interpretation**: Assess the reliability of results and get conclusions.

19.6 Example: The Impact of Apple Origins on Purchasing Prices

Structural Causal Model

Case study

In agricultural economic research, identifying the true effect of product origin on market prices is crucial for guiding production layout and optimizing resource allocation. However, this causal inference is often complicated by multiple non-independent influencing factors, such as the intrinsic quality attributes of agricultural products. This example focuses on the causal effect of *apple origin* (X) on *purchase price* (Y). Other influencing factors include sweetness (A), apple size (B), and shelf life (C), where A, B, and C are not mutually independent. Table 19.6 is the price of apples for different variable values.

Variables	(12,4,3)	(12,4,7)	(12,6,3)	(12,6,7)	(14,4,3)	(14,4,7)	(14,6,3)	(14,6,7)
Price 1	¥5	¥6	¥9	¥10	¥7	¥9	¥10	¥16
Price 2	¥11	¥13	¥16	¥16	¥13	¥16	¥18	¥19
Rarity 1	0.25	0.25	0.2	0.1	0.05	0.05	0.05	0.05
Rarity 2	0.2	0.2	0.2	0.15	0.1	0.05	0.03	0.07

Table 19.1: The price (in Yuan(¥)) and rarity (denoted with probability weight) of apples from two origins. "Variables" stands for (Brix, kg, days).

- *Endogenous Variables*: Variables influenced by other variables within the model.
  - X (Apple Origin): 1 = target origin, 0 = other origins;
  - Y (Apple Purchase Price): Continuous variable (unit: e.g., yuan/kg);
  - A (Apple Sweetness): Continuous variable (unit: e.g., Brix value);
  - B (Apple Size): Continuous variable (unit: e.g., diameter in mm or weight in g per fruit);
  - C (Apple Shelf Life): Continuous variable (unit: e.g., days).

- *Exogenous Variables*: Variables not influenced by model-internal variables, serving only as inputs.
  - W_X (Unobserved factors for X): e.g., trace elements in soil of the origin, climate stability;
  - W_Y (Unobserved factors for Y): e.g., short-term capital pressure of buyers, sudden market demand;
  - W_A (Unobserved factors for A): e.g., genetic differences of apple tree varieties, sunlight duration during growth;
  - W_B (Unobserved factors for B): e.g., rainfall during growth, uniformity of fertilization;
  - W_C (Unobserved factors for C): e.g., preliminary processing after picking, storage conditions before transportation.

- *Structural Functions (F)*: Define causal relationships between endogenous variables, reflecting non-independence among A, B, and C.

$$
\begin{aligned}
X &:= f_X(W_X) \\
A &:= f_A(X, W_A) \\
B &:= f_B(X, A, W_B) \\
C &:= f_C(A, B, W_C) \\
Y &:= f_Y(X, A, B, C, W_Y)
\end{aligned}
$$

The causal relationships between multiple variables in this example is presented in Figure 19.1.

Calculation

- *Front-Door Paths*: Including all paths from X to Y. The causal flows between them include: $X \to A \to B \to C \to Y$; $X \to B$; $A \to C$; $X \to Y$; $A \to Y$; $B \to Y$. All exogenous variables W point to their corresponding endogenous variables (e.g., $W_A \to A$).

- *Condition Satisfaction*: Compute with all concerning variables $\{A, B, C\}$ (sweetness, size, shelf life).

1. There are no back-door paths from X to Y, therefore non-causal associations are not involved;
2. There are no ancestors of X in endogenous variables (ancestors refer to variables that can influence X. A, B, and C are influenced by X, not influencing X);
3. Thus, the interventional distribution $\mathbb{P}(Y = y \mid \mathbf{do}(X = x))$ is identifiable.

- *Interventional Distribution*: Calculate the distribution of purchase price under a specific origin ($\mathbf{do}(X = x)$) with formula 19.5 and without cutting-off the paths between X and Y:

$$\begin{aligned}&\mathbb{P}(Y = y \mid \mathbf{do}(X = x))\\ &= \sum_{a,b,c} \mathbb{P}(Y = y \mid X = x, A = a, B = b, C = c)\end{aligned}$$

  Since A, B, and C are not independent, the joint probability $\mathbb{P}(A, B, C)$ is used instead of the product of their marginal probabilities.

- *Average Treatment Effect (ATE)*: Measures the average causal effect difference of "target origin" vs. "other origins" on purchase price:

$$\text{ATE} = \mathbb{E}[Y \mid \mathbf{do}(X = 1)] - \mathbb{E}[Y \mid \mathbf{do}(X = 0)]$$

- *Calculation*: Substitute the concrete values into ATE to get a computable expression:

$$\begin{aligned}&\text{ATE} =\\ &\sum_{a,b,c} \mathbb{E}[Y \mid X = 1, A = a, B = b, C = c]\\ &- \sum_{a,b,c} \mathbb{E}[Y \mid X = 0, A = a, B = b, C = c]\\ &= (5 \times 0.25 + 6 \times 0.25 + 9 \times 0.2 + 10 \times 0.1 + 7 \times 0.05 + 9 \times 0.05\\ &+ 10 \times 0.05 + 16 \times 0.05) - (11 \times 0.2 + 13 \times 0.2 + 16 \times 0.2 + 16 \times 0.15\\ &+ 13 \times 0.1 + 16 \times 0.05 + 18 \times 0.03 + 19 \times 0.07)\\ &= -6.72\end{aligned}$$

Conclusion

The ATE quantifies the average causal effect of *apple origin* on *purchase price*. By computing with the mediator variables (sweetness A, size B, and shelf life C), this effect excludes the interference of confounding factors and only reflects the true impact of "different origins" on the purchase price. The result reports apples grown in Origin 1 had an average purchasing price of ¥6.72/kg *lower* than that of apples grown in Origin 2. Therefore, it is confident that this difference is likely due to the origin itself and not other confounding factors.

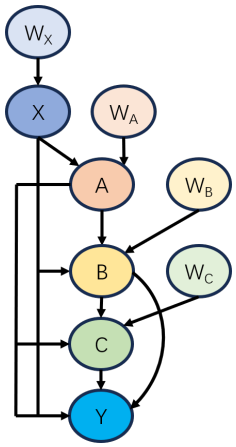


Figure 19.1: The causal relationship model between multiple variables in this example.

19.7 Limitations

SCM is a robust framework for understanding and analyzing causal relationships. However, SCM has the following limitations.

Firstly, it completely relies on the correctness of the causal assumptions as a DAG. If the assumptions are incorrect or not confident, the resulting conclusions may be invalid or misleading.

Secondly, SCM also requires sufficient background knowledge to justify the causal structure, as the evaluation results of SCM cannot specify the true causal-effect relationship.

Thirdly, SCM often necessitates large datasets, particularly in complex models with many variables. Measurement error or unobserved confounding can further compromise the validity of SCM results, as the framework assumes that all relevant variables are correctly measured and included.

Lastly, the computation will be cost-consuming or even infeasible when dealing with high-dimensional data or when attempting to estimate effects in a complicated DAG.

19.8 Summary

Structural causal model is a comprehensive framework for representing and analyzing causal effects among variables. SCM extends traditional statistical association analysis by introducing a framework that represents causal relationships through a system of structural equations and a corresponding graphical model. This framework is applied in fields such as epidemiology, economics, computer science, and social sciences for tasks including causal discovery, mediation analysis, and evaluation of intervention effects. The limitations of SCM is the high cost of data collecting and the unreliable assumptions and causal diagrams.

Chapter 20

Structural Equation Model

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the Structural Equation Model (in short, SEM).

20.1 Basic Concepts

Observed Variables are the variables whose values can be directly obtained in a study [96], such as survey responses, test scores, or any other quantifiable data points. In SEM, they serve as the observed indicators, that is, the observed variables at the statistical level, for latent variables, and their sample covariance matrix provides the empirical basis for a model estimation.

Latent Variables are unobserved variables that can only be inferred indirectly from a theoretical model [96]. They usually represent unobservable constructs, such as psychological traits (e.g., motivation, intelligence) or sociological concepts (e.g., social influence, group norms).

Factor. In SEM, the factor represents the latent variable at the statistical level and is identified through its systematic relationships with observed indicators [96].

Exogenous Variables are "given from the outside [96]," and serve as the theoretical model's inputs or causes. They can be *observed or latent variables* and may co-vary with each other. They influence endogenous variables.

Endogenous Variables are "accounted for by the model [96]." Their variance is explained by other variables within the model. They can be *observed or latent variables.*

Factor Analysis aims to describe the covariance relationships among many variables in terms of a few underlying factors. [95]

A Measurement Model is a model linking latent variables to their observed indicators [21]. It specifies how each latent variable is operationally defined and measured. The measurement model in SEM is essentially a confirmatory factor analysis that specifies how observed variables reflect their underlying latent variables.

[1] The primary contributor is Dr. Lei Wang.

Confirmatory Factor Analysis (CFA) "deals specifically with measurement models—that is, the relationships between observed measures or indicators, e.g., test items, test scores, behavioral observation ratings, and latent variables or factors [24]."

Path Analysis is a statistical method that estimates effects among a set of observed variables based on a hypothesized causal structure, which is visually represented in a path diagram where unidirectional arrows denote the theorized directional influences [218].

A Structural Model specifies the hypothesized causal relationships among variables, particularly between latent variables [96]. In this context, causal relationships refer to the directional influences where changes in an independent variable are theorized to produce changes in a dependent variable. A structural model is a broader conceptual framework that generalizes and extends the principles of path analysis.

A Theoretical Model is a formal specification of hypothesized relationships, which is defined by a measurement model and a structural model.

20.2 Problem Statement

SEM [96, 207, 87, 5, 101] is a powerful statistical technique that allows researchers to analyze complex relationships between variables. SEM is widely used in fields requiring multivariate analysis, especially in the behavioral sciences, where it is used to model psychological, social, and economic constructs.

SEM is a general method for estimating the unknown coefficients in a set of linear structural equations [96]. The SEM problem could be stated as "how to create and validate a theoretical model that includes multiple abstract concepts (latent variables) and their complex causal relationships based on observable data." SEM operationalizes the theoretical model by estimating parameters to quantify the direction and strength of influences, illustrating how exogenous variables affect endogenous variables. Specifically, it simultaneously addresses two questions: "How to accurately measure abstract concepts?" and "What are the causal network relationships between these concepts?"

This problem can be further formulated as a detailed analysis problem. It begins with two primary inputs: the empirical data, which consist of measurements of the observed variables (e.g., survey responses or test scores), and the theoretical model specification. The theoretical model specification articulates the hypothesized relationships using the fundamental concepts of SEM: it defines the latent variables (unobservable constructs) and links them to their observed variables (indicators), while also positing the causal network between exogenous variables (the external inputs or causes, which can be observed or latent) and endogenous variables (the outcomes whose variance is explained by the model). The model is subject to five key constraints, as detailed in Section 20.3. Then, the core task of SEM is to estimate the unknown coefficients in this system of linear structural equations, testing how well the specified model explains the observed data.

20.3 Basic Assumptions

Assumption 20.1 (Independence of Measurement Errors)**.** *The error terms associated with the observed indicators have an expected value of zero, are uncorrelated with each other, and are uncorrelated with the latent constructs they are intended to measure.*

Assumption 20.2 (Multivariate Normal Distribution)**.** *The observed variables exhibit a multivariate normal distribution.*

Assumption 20.3 (Linearity)**.** *SEM typically assumes linear relationships across its components, including linear associations between latent variables and their observed variables, as well as between latent variables themselves.*

Assumption 20.4 (Independence of Structural Residuals and Exogenous Variables)**.** *The structural disturbance terms (errors in equations) have an expected value of zero and are uncorrelated with the exogenous latent variables in the model.*

Assumption 20.5 (Model Identification)**.** *The model must be identified, meaning that there is a unique set of parameter values consistent with the population covariance matrix of the observed variables.*

20.4 Basic Principles

Structural Equation Modeling (SEM) is a statistical framework that combines *Confirmatory Factor Analysis (CFA)* and *Path Analysis* to test theoretical models by evaluating the fit between the covariance structure predicted by the model is represented by its covariance matrix. and the covariance structure observed in empirical data. SEM allows for the simultaneous estimation of both measurement and structural components, providing a comprehensive view of the relationships between latent and observed variables.

A Measurement Model defines how latent variables are measured by observed indicators. CFA is used to validate the hypothesis that a set of observed variables represents an underlying concept. The *Structural Model* specifies the causal relationships between latent variables or between latent and observed variables. The structural models direct and indirect effects in a way that is similar to a system of simultaneous regression equations, allowing for complex interdependencies.

A key strength of SEM is its ability to account for measurement error, unlike traditional regression models that assume independent variables are measured without error. SEM incorporates error terms for both latent and observed variables, resulting in more accurate and unbiased estimates.

The parameters of the SEM include factor loadings (coefficients that measure the strength and direction of a latent variable's influence on its observed indicators), path coefficients (coefficients that represent the strength and direction of the direct influence of one variable on another), variances (primarily including the variances of exogenous

variables, residuals, and measurement errors), and covariances (mainly referring to the associations between exogenous variables, which are often standardized into correlation coefficients for interpretation).

To estimate these parameters, SEM relies primarily on statistical estimation methods that compare the hypothesized model with the observed covariance matrix of the data. The most commonly used method is *Maximum Likelihood (ML)* estimation, which seeks parameter values that maximize the likelihood of observing the data given the assumed model structure. The fit between the hypothesized model and observed data is tested using statistical indices such as the chi-square test (χ^2), Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR).

20.5 Methodology

The SEM methodology involves a process, beginning with the formal specification of the model, followed by parameter estimation, and concluding with model evaluation.

SEM is formally represented by a set of linear equations, which are typically divided into two components:

Measurement Model. The measurement component specifies how the latent variables are reflected in the observable indicators. Formally, the relationships for exogenous and endogenous measurements are given by:

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \delta, \tag{20.1}$$

$$\mathbf{y} = \Lambda_y \boldsymbol{\eta} + \epsilon, \tag{20.2}$$

where $\mathbf{x}$ denotes the observed indicators associated with the exogenous latent variables $\boldsymbol{\xi}$, and $\mathbf{y}$ denotes the observed indicators associated with the endogenous latent variables $\boldsymbol{\eta}$. The matrices Λ_x and Λ_y represent the corresponding factor loadings, while δ and ϵ capture measurement errors for $\mathbf{x}$ and $\mathbf{y}$, respectively.

Structural Model. The structural component characterizes the causal relations among the latent variables. Specifically, the endogenous latent variables are determined by both other endogenous factors and the exogenous inputs:

$$\boldsymbol{\eta} = B\boldsymbol{\eta} + \Gamma\boldsymbol{\xi} + \zeta, \tag{20.3}$$

where $\boldsymbol{\eta}$ represents the vector of endogenous latent variables and $\boldsymbol{\xi}$ denotes the exogenous latent variables. The matrix B captures the regression relations among endogenous variables, whereas Γ represents the effects of exogenous variables on endogenous ones. The term ζ accounts for structural residuals(disturbances) not explained by the model.

These equations form the foundation of SEM, enabling the simultaneous modeling of the measurement structure of latent constructs and the structural (causal or associative) relationships among them.

The primary method for estimating parameters in SEM is *Maximum Likelihood (ML) estimation*, which assumes that the observed variables follow a *multivariate normal distribution*. ML estimates model parameters—such as factor loadings, path coefficients, variances, and covariances—by minimizing the difference between the observed sample covariance matrix and the covariance matrix implied by the hypothesized model. Under the assumption of multivariate normality, ML provides efficient and unbiased parameter estimates.

The ML estimation method also generates key *goodness-of-fit indices* that help test how well the model reproduces the observed data structure. After estimating the model, researchers evaluate its overall fit using a combination of *goodness-of-fit indices*. While the *Chi-square (χ^2) statistic* tests the exact fit between the observed and model-implied covariance matrices, it is highly sensitive to sample size and is therefore interpreted cautiously. More practically, researchers rely on *absolute fit indices* (e.g., *RMSEA*, *SRMR*) and *incremental fit indices* (e.g., *CFI*) to test model adequacy. Commonly used fit criteria include:

- *CFI (Comparative Fit Index)*: Values greater than 0.90 are generally regarded as indicative of an acceptable model fit, whereas values exceeding 0.95 denote a good fit.

- *RMSEA (Root Mean Square Error of Approximation)*: Values below 0.08 typically indicate an acceptable fit, while those below 0.05 reflect a close or good fit.

- *SRMR (Standardized Root Mean Square Residual)*: Values below 0.08 are commonly interpreted as indicative of an acceptable fit.

These indices should be interpreted in combination, taking into account model complexity, theoretical expectations, and sample characteristics.

20.6 Example: Calculate the Effects of Apple Origins on Purchasing Prices

This example uses a hybrid SEM approach to explore how *the apple origins* influence the purchasing price through key quality characteristics. We employ a mixed-variable design where an origin is modeled as a latent construct, while apple characteristics and prices are directly observed. This approach balances methodological rigor with practical interpretability.

Structural Equation Model

Variable Specification

The model uses a combination of latent and manifest variables based on agricultural measurement standards:

- *Origin (OR)*: *Latent variable* representing regional cultivation advantage, measured by 3 observed indicators:
  - OR1: Climate suitability (1-5 scale, 5 = most suitable for apple growth)
  - OR2: Soil organic matter content (g/kg, 15-25 g/kg = optimal range)
  - OR3: Average annual temperature (°C, 8-12 °C = ideal for sugar accumulation)
- *Sweetness (SW)*: *Observed variable* measured in Brix (continuous):
  - Typical range: 10-16 Brix for commercial apples
- *Size (SI)*: *Observed variable* (continuous):
  - Typical range: 120-220 g for commercial apples
- *Shelf Life (SL)*: *Observed variable* measured in days (continuous):
  - Typical range: 3-14 days under normal temperature storage
- *Purchasing Price (PP)*: *Observed variable* measured in ¥/kg:
  - Directly observed market transaction price

Note: In the path model, SW, SI, SL, and PP function as endogenous variables (being influenced by other variables), while OR serves as the exogenous latent variable.

Theoretical Framework and Hypotheses

Based on agricultural economics research, we hypothesize that superior origin conditions directly enhance apple quality attributes, which in turn increase market value. The proposed causal pathways are:

$$OR \to SW, \quad OR \to SI, \quad OR \to SL$$
$$SW \to PP, \quad SI \to PP, \quad SL \to PP$$

Theoretical Rationale:

- Better origin conditions (optimal climate, soil, temperature) promote sugar accumulation (↑ sweetness), fruit development (↑ size), and structural integrity (↑ shelf life)
- Higher quality attributes command price premiums in the market due to consumer preferences for sweeter, larger, and longer-lasting apples
- Origin may also exert a direct price effect through regional branding and reputation effects

Model Specification

1. the Measurement Model (for the only exogenous latent variable *Origin*):

$$OR1 = \lambda_{OR1} OR + \delta_{OR1}, \quad (\lambda_{OR1} = 1, \text{ marker variable for identification})$$
$$OR2 = \lambda_{OR2} OR + \delta_{OR2}$$
$$OR3 = \lambda_{OR3} OR + \delta_{OR3}$$

where λ = factor loading, δ = measurement error. The outcome of the measurement model—the factor scores of the latent variable Origin (OR)—is computed and subsequently treated as observable data within the structural model for the estimation of path coefficients and future predictions.
2. the Structural Model (path analysis among all variables):

$$SW = \beta_1 OR + \zeta_1 \qquad \text{(Origin} \rightarrow \text{Sweetness)} \tag{20.4}$$

$$SI = \beta_2 OR + \zeta_2 \qquad \text{(Origin} \rightarrow \text{Size)} \tag{20.5}$$

$$SL = \beta_3 OR + \zeta_3 \qquad \text{(Origin} \rightarrow \text{Shelf Life)} \tag{20.6}$$

$$PP = \gamma_1 OR + \gamma_2 SI + \gamma_3 SW + \gamma_4 SL + \zeta_4 \qquad \text{(Total effects} \rightarrow \text{Price)} \tag{20.7}$$

where β, γ = path coefficients, ζ = structural disturbances.

Parameter Estimation and Model Identification

Estimation Results (hypothetical but realistic parameters based on agricultural research):

- **Measurement model parameters:**

$$\begin{aligned} \lambda_{OR1} &= 1.00 \quad \text{(fixed for identification)} \\ \lambda_{OR2} = 0.75, &\quad \lambda_{OR3} = 0.68 \quad \text{(freely estimated)} \\ \text{Var}(\delta_{OR1}) = 0.20, &\quad \text{Var}(\delta_{OR2}) = 0.25, \quad \text{Var}(\delta_{OR3}) = 0.30 \end{aligned}$$

- **Structural model parameters:**

$$\begin{aligned} \beta_1 = 0.55, &\quad \beta_2 = 0.48, \quad \beta_3 = 0.62 \\ \gamma_1 = 0.25, &\quad \gamma_2 = 0.30, \quad \gamma_3 = 0.35, \quad \gamma_4 = 0.20 \end{aligned}$$

All parameters are statistically significant at $p < 0.05$, supporting the hypothesized relationships.

Numerical Prediction Example

This example demonstrates how the model can be used for prediction. With estimated parameters, the structural equation for price is:

$$\hat{PP} = 0.25 \times OR + 0.30 \times SI + 0.35 \times SW + 0.20 \times SL + \zeta_4$$

For apples with specific quality attributes:

$$SW = 14 \text{ Brix}, \quad SI = 165g, \quad SL = 7 \text{ days}$$

For a region with above-average origin conditions (where the origin condition score, OR = 3.8, is calculated as a weighted combination of the region's OR1, OR2, and OR3 values, with the weights determined by the measurement model parameters), the predicted price component derived from observed factors is:

$$\begin{aligned} \hat{PP} &= 0.25 \times 3.8 + 0.30 \times 1 + 0.35 \times 14 + 0.20 \times 7 \\ &= 0.95 + 0.30 + 4.90 + 1.40 = 7.55 \text{ ¥/kg} \end{aligned}$$

Accounting for unmodeled factors ($\zeta_4 = -1.05$), the final predicted price is:

$$\hat{PP} = 7.55 - 1.05 = 6.50 \text{ ¥/kg}$$

Effect Decomposition Analysis

We decompose the total effect of the apple origins on the purchasing price to quantify mediation pathways.

1. Indirect effects:

$$\begin{aligned} \text{Through sweetness:} &\quad \beta_1 \times \gamma_3 = 0.55 \times 0.35 = 0.1925 \\ \text{Through size:} &\quad \beta_2 \times \gamma_2 = 0.48 \times 0.30 = 0.1440 \\ \text{Through shelf life:} &\quad \beta_3 \times \gamma_4 = 0.62 \times 0.20 = 0.1240 \\ \text{Total indirect effect:} &\quad 0.1925 + 0.1440 + 0.1240 = 0.4605 \end{aligned}$$

2. Direct effects :

$$\text{Direct effect} = \gamma_1 = 0.25$$

3. Total effect:

$$\text{Total effect} = \text{Direct} + \text{Indirect} = 0.25 + 0.4605 = 0.7105$$

Based on the standardized results, the *total effect* of origin advantage (OR) on price (PP) is 0.7105. This means that when OR increases by one standard deviation, PP increases by 0.71 standard deviations on average. The *direct effect* is 0.25, and the *indirect effect* is 0.4605—mediated through sweetness (0.1925), size (0.1440), and shelf life (0.1240)—corresponding to a *direct share* of 35.2% and an *indirect share* of 64.8%.Figure 20.1 is the graphical representation.

Model Fit and Validation

Goodness-of-Fit Indices (hypothetical results):

- $\chi^2/df = 2.15$ (acceptable: < 3.0)
- CFI = 0.94 (good fit: > 0.90)
- RMSEA = 0.06 (good fit: < 0.08)
- SRMR = 0.05 (good fit: < 0.08)

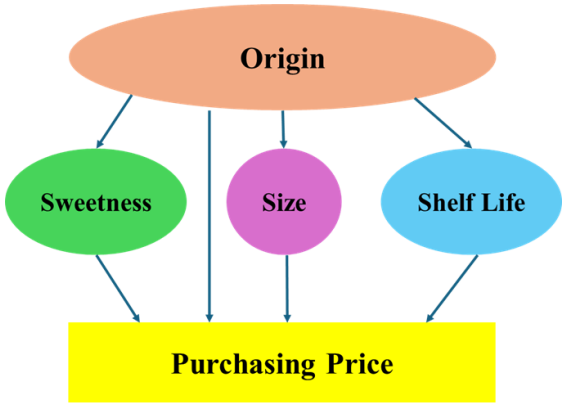


Figure 20.1: SEM path diagram for apple pricing model

In summary, this example demonstrates SEM's applicability in agricultural price analysis—by standardizing indicators (10-16 Brix sweetness, 120-220g size, 3-14 day shelf life), the model's results are more actionable for producers and marketers. The impact decomposition clarifies that the apple origin influences the purchasing price mainly through improving sweetness. This insight can help regions focus on sugar accumulation techniques to enhance apple market value.

20.7 Limitations

SEM is a powerful tool. However, SEM has limitations that researchers must consider. It generally requires many assumptions to be satisfied, including linearity, normality, and independence of errors. It also generally requires a large sample size, especially for complex models, as small samples can lead to unreliable results. Model misspecification can also produce misleading conclusions, emphasizing the importance of grounding model design in theoretical and empirical evidence. Furthermore, SEM can be computationally intensive, particularly when incorporating nonlinear relationships or categorical variables, necessitating robust software and computational resources. Lastly, while SEM can identify associations between variables, it does not inherently establish causality, so causal inferences should be made cautiously.

20.8 Summary

Structural Equation Model (SEM) is a powerful tool used to analyze complex relationships in multivariate data. It allows researchers to model latent variables, handle measurement errors, and examine the structure of relationships across various fields, such as psychology, sociology, education, and marketing. SEM is particularly useful for understanding how multiple variables interact simultaneously, making it invaluable in the behavioral sciences. However, it is essential to pay close attention to model spec-

ification, data quality, and sample size to avoid common pitfalls such as overfitting or misspecification.

Chapter 21

The Potential Outcome Theory

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and a case study of the potential outcome (PO) theory.

21.1 Basic Concepts

A Unit refers to "physical objects at particular points in time, e.g., plots of land, individual people, one person at repeated points in time" [158].

A Treatment represents the intervention or condition assigned to a unit, often denoted by a binary indicator T, where $T = 1$ denotes the active treatment and $T = 0$ the control treatment [158]. Generally, PO uses treatment to denote "active treatment" and control to denote "control treatment," respectively [158].

An Assignment Mechanism refers to "a probabilistic model for the treatment each unit receives as a function of covariates and potential outcomes" [158].

A Covariate represents "a variable that takes its values before the treatment assignment or cannot be affected by the treatment, such as the sex of the unit [158]." Covariates are also called pretreatment variables [155].

A Posttreatment Variable refers to "the variables describing experimental units that might be recorded after the assignment of treatment [155]." These variables may be influenced by the treatment itself or by other post-assignment factors. For example, in an agricultural experiment where the treatment is the use of a specific fertilizer, the *chlorophyll content* or *fruit size* measured at harvest would be posttreatment variables, since they are recorded after the fertilizer has been applied [155].

A Mediator is a specific type of posttreatment variable that not only occurs after the treatment but also transmits part of the treatment's causal effect to the outcome [157]. Formally, for each unit i, let $M_i(t)$ denote the potential outcome of the mediator under treatment level t, and let $Y_i(t, m)$ denote the potential outcome as a function of both the treatment t and the mediator m. Under this framework, the *natural direct effect (NDE)*

[1] The primary contributor is Dr. Wanling Gao.

and *natural indirect effect (NIE)* are defined as [140, 157]:

$$NDE = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))], \tag{21.1}$$

$$NIE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))]. \tag{21.2}$$

The mediator thus represents a causal pathway through which the treatment influences the outcome, beyond the direct effect of treatment itself.

PO refers to the outcome that would be observed for a unit at a particular point in time after a specific action (i.e., treatment or control). Specifically, $Y(1)$ represents the PO under the active treatment and $Y(0)$ under the control treatment. Only one of these outcomes can be observed in reality, depending on the treatment actually received [158].

A Counterfactual refers to the unobserved potential outcome for a unit under a treatment condition different from the one it actually received. Since each unit can experience only one treatment state, the counterfactual outcome remains unobserved and represents the core challenge of causal inference, estimating what would have happened under the alternative treatment scenario.

Causal Effects are defined as "comparisons of POs under different treatments on a common set of units [158]."

21.2 Problem Statement

This section first illustrates the fundamental problem of causal inference under the PO framework and then describes its problem formulation.

21.2.1 Fundamental Problem of Causal Inference: Missing Data Problem

A crucial implication of this framework is that causal inference is inherently a missing data problem [89]: for each unit, at most one of the potential outcomes is observed, while the other remains counterfactual. This creates the central challenge of causal inference —how to estimate treatment effects when one of the outcomes is always unobserved.

More fundamentally, at any given time point, a unit cannot simultaneously both receive and not receive a treatment, nor can it be exposed to multiple alternative treatments at once. Only one realized path is observed, while all other potential outcomes remain unobserved counterfactuals. Although a unit can indeed receive different treatments across different time periods, the states of the unit are not interchangeable across time. Outcomes at later time points may be influenced not only by the treatment assigned at that time, but also by earlier interventions or by evolving characteristics of the unit itself. For example, if a patient takes a drug at time t, the observed effect at time $t+1$ may reflect not only the drug's direct causal impact, but also prior exposures, accumulated biological responses, or progression of the underlying disease. These time-varying factors act as confounders, complicating the attribution of observed differences solely to the treatment of interest.

21.2.2 Problem Statement

Given a population of observational or experimental units, each of which can receive one of several treatment conditions ($T \in \{0, 1\}$), let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes that unit i would exhibit under the treatment and control conditions, respectively. For each unit, only one potential outcome can be observed, depending on the treatment actually assigned, while the other remains unobserved.

The unknown quantity is the unobserved potential outcome for each unit, which prevents direct observation of the individual-level treatment effect ($Y_i(1) - Y_i(0)$). This limitation defines the core challenge of the potential outcome framework: the impossibility of simultaneously observing both potential outcomes for the same unit.

The PO problem is stated as "how to infer or estimate the causal effect of the treatment on the outcome based on observed data comprising the treatment assignment, covariates, and realized outcomes." It begins with one primary input: the observed data, consisting of the realized outcome for each unit together with its treatment assignment and pre-treatment covariates. The model is subject to a set of well-established constraints as detailed in Section 21.3 that govern the relationship between the potential outcomes, the assignment mechanism, and the observed data. Given these ingredients, the core task of causal inference under the potential outcomes framework is to identify and estimate causal estimands—such as the average treatment effect (ATE) or related population-level contrasts—using only the observed data together with the assumptions encoded in the PO. The framework thereby provides a disciplined way to formalize what aspects of the causal effect are identifiable and to what extent the observed data support inferences about the unobserved potential outcomes.

Inputs. The observed data $\mathcal{D}$ consist of

$$\mathcal{D} = \{(Y_i^{\text{obs}}, T_i, X_i)\}_{i=1}^n, \tag{21.3}$$

where $T_i \in \{0, 1\}$ is the treatment assignment and X_i denotes pre-treatment covariates.

Outputs (Causal Estimands). The target quantities are causal effects defined in terms of the joint distribution of the potential outcomes, including:

$$\text{ATE: } \mathbb{E}[Y_i(1) - Y_i(0)], \tag{21.4}$$

$$\text{ATT: } \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1], \tag{21.5}$$

and covariate-conditional effects such as the CATE,

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{21.6}$$

Constraints (Assumptions). The PO is subject to three key constraints, as detailed in Section 21.3.

21.3 Basic Assumptions

To make causal effects identifiable and estimable, the PO framework relies on three foundational assumptions:

Assumption 21.1 (Stable Unit Treatment Value Assumption (SUTVA) [43, 158].)**.** *A fundamental assumption in causal inference is SUTVA. Each unit's PO depends only on its own treatment status; there is no interference between units [43, 158], and treatments are well-defined. This principle ensures that PO is well-defined and comparable across units. It consists of two essential components: (1) no interference between units, and (2) no hidden versions of treatments [158].*

First, the no-interference condition requires that the treatment assigned to one unit does not alter the PO of another. Consider, for example, the evaluation of a personalized recommendation system in an online learning platform. If we assume that the courses recommended to Student A do not influence Student B's learning outcomes, then this part of SUTVA is satisfied. In practice, however, this assumption may fail. Suppose the platform has limited seats in certain courses; if Student A is recommended and enrolls in a popular class, Student B may be excluded from it, thereby indirectly affecting B's outcome. In such settings, interference needs to be explicitly addressed to avoid biased causal estimates.

Second, the assumption that there is no hidden version of treatments requires that each treatment level be uniquely and consistently defined. If we define the treatment as "participating in a training program," it is crucial that this program is homogeneous across participants. If some trainees attend a three-day workshop while others undergo a three-month intensive curriculum, then the label "training" actually corresponds to multiple distinct interventions. In this case, potential outcomes cannot be uniquely mapped to a single treatment level, violating SUTVA. A common strategy to resolve this issue is to refine the definition of treatment—distinguishing, for instance, between "short-term training" and "long-term training," so that each POS corresponds clearly to a well-defined intervention.

In summary, SUTVA provides the conceptual clarity needed for defining and estimating causal effects. Without it, the PO may be confounded by interference across units or ambiguities in treatment definitions. Although in many applied settings SUTVA may only hold approximately, recognizing its role and limitations is an indispensable step in credible causal inference.

Assumption 21.2 (Ignorability (Unconfoundedness) [151, 154])**.** *Conditional on observed covariates X, treatment assignment is independent of potential outcomes $(Y(0), Y(1))$:*

$$(Y(0), Y(1)) \perp\!\!\!\perp T \mid X. \tag{21.7}$$

This assumption is the cornerstone for establishing causality from observational data. The symbol "$\mid$" is a mathematical notation denoting "under the condition of," and "$A \mid B$"

signifies "A under the condition of B." The symbol "⫫" represents "mutual independence," with "$A \perp\!\!\!\perp B$" indicating that "A and B are mutually independent." The formula asserts that, once we account for (or condition on) all relevant pre-treatment covariates X, the mechanism that determines whether a unit receives the treatment ($T = 1$) or control ($T = 0$) becomes independent of what the outcomes would have been under either scenario. In essence, after controlling for X, the treatment groups are comparable, and any remaining difference in outcomes can be attributed to the treatment itself. This is equivalent to saying there are no unmeasured confounders: all common causes of T and Y are captured in X.

Consider an online learning platform assessing the effect of a new premium interactive module ($T = 1$) versus the standard module ($T = 0$) on final exam scores (Y), where students self-select into the premium module. The ignorability assumption requires that we can measure all covariates X (e.g., prior GPA, time spent on the platform, past quiz scores) that influence both a student's decision to enroll and their eventual exam score. If this assumption holds, then among students who are identical on these measured covariates X, choosing the premium module is as good as a random assignment. This allows for a fair comparison, as the groups are balanced with respect to both observed and, critically, any unobserved factors that are correlated with the observed X.

Assumption 21.3 (Positivity (Overlap) [151])**.** *For all possible values of covariates X, the probability of treatment assignment is strictly between 0 and 1:*

$$0 < P(T = 1 \mid X) < 1. \tag{21.8}$$

This ensures sufficient overlap in covariate distributions between treatment and control groups. In the online learning platform example, this assumption requires that for every student profile (e.g., all ranges of prior achievement levels), there exists a non-zero probability of both enrolling and not enrolling in the new module. If the platform restricts module access based on certain criteria (e.g., prohibiting low-performing students from enrolling), the assumption is violated for the affected subpopulations, rendering causal inference infeasible in those strata due to a lack of counterfactual data.

These assumptions provide the theoretical scaffolding that allows researchers to link observed or experimental data to causal effects.

21.4 Origin of Ideas

PO, also known as the Neyman–Rubin Causal Model (NRCM), is one of the most influential paradigms in modern causal inference. Its intellectual roots trace back to Jerzy Neyman (1923) [183, 156], who first introduced the concept of PO in the context of randomized experiments in agricultural research.

In his 1923 doctoral thesis, Jerzy Neyman introduced the notion of "potential yields" to analyze agricultural field experiments. He imagined that for each plot of land, we could think of the potential yield that would result if that plot were planted with each of

the different varieties. In reality, however, only one crop variety can actually be planted on a given plot, so we only get to observe one of these potential yields. The other yields remain unobserved. This way of thinking —that each unit (in this case, each plot of land) has multiple possible outcomes, only one of which is realized depending on the treatment it receives (which crop is planted) —is the basic idea behind the modern PO framework.

Neyman then asked a practical question: if we want to compare the average yields of two varieties, how do we calculate the uncertainty (variance) of that comparison, given that we cannot observe all the potential yields?

Neyman derived a formula for the variance of the difference between the observed averages. This variance depends not only on the variability of yields across plots but also on the correlation r between the potential yields of the two varieties on the same plot. Here, r represents how similarly the two varieties would have performed on the same piece of land if each had been planted. The challenge is that r cannot be directly observed—since a plot can host only one variety—so it introduces uncertainty into the variance estimate.

To address this uncertainty, Neyman adopted a conservative strategy, setting $r = 1$. While this assumption may appear counterintuitive, it functions as a *principled lower-bound estimate* for the variance. Specifically, the variance of the difference between the sample means of two varieties can be expressed as

$$\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \frac{\sigma_1^2}{p_1} + \frac{\sigma_2^2}{p_2} - 2r\frac{\sigma_1\sigma_2}{p}, \tag{21.9}$$

where σ_1^2 and σ_2^2 are the variances of the potential yields for the two varieties, r is the correlation between their potential yields on the same plot, and p is the number of plots. The corresponding standard error (SE) is given by

$$\mathrm{SE} = \sqrt{\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2)}, \tag{21.10}$$

and the t-statistic for testing the difference in means is

$$t = \frac{\text{Estimated Difference}}{\mathrm{SE}}. \tag{21.11}$$

Since r is unobservable, setting $r = 1$ maximizes the covariance term $2r\sigma_1\sigma_2/n$, which in turn *reduces the subtracted portion in the variance formula* and thus yields a *larger estimated variance*. A larger variance increases the standard error, producing a more conservative t-statistic. This ensures that the comparison does not overstate significance: if the observed difference is non-significant, one can be reasonably confident that the lack of effect is genuine, minimizing the risk of false-positive conclusions. Conversely, any significant finding should still be interpreted with caution, acknowledging that the true correlation may differ from the assumed upper bound.

This idea of "potential yield" was Neyman's original way of formalizing the fact that outcomes depend on the treatment assigned. Decades later, Donald Rubin (1972 [154],

1974, 1978 [155], 2004 [157], 2005 [158], 2006 [159]) reformulated and generalized this concept under the name PO, which has since become the foundation of modern causal inference. Rubin's contribution transformed Neyman's idea into a general statistical methodology for causal reasoning, applicable well beyond randomized trials.

21.5 Basic Principles

Rubin elegantly reframes this causal inference issue as a systematic attempt to recover or approximate these missing counterfactuals through careful design, appropriate modeling, or both.

In Rubin's formalization, each unit i in a study is associated with a pair of potential outcomes: $(Y_i(1), Y_i(0))$, where $Y_i(1)$ represents the outcome if the unit receives the treatment, and $Y_i(0)$ represents the outcome if the unit does not. The observed outcome is given by:

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0), \tag{21.12}$$

where $T_i \in \{0, 1\}$ is the treatment assignment indicator.

This representation makes explicit the fact that only one of the two outcomes can ever be observed for each unit. The causal effect at the individual level is thus defined as the *Individual Treatment Effect*:

$$\tau_i = Y_i(1) - Y_i(0), \tag{21.13}$$

while population-level effects, such as the *Average Treatment Effect (ATE)*, are defined as:

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{21.14}$$

Since individual effects are generally unidentifiable, most analyses focus on population-level averages such as the ATE, or subgroup-level effects like the Conditional Average Treatment Effect (CATE).

21.6 Methodology

Within the PO framework, the estimation of causal effects proceeds from the fundamental problem of missing counterfactuals—only one of $Y_i(1)$ or $Y_i(0)$ is observable for each unit i. Therefore, the methodological focus lies in constructing valid comparisons that approximate the distribution of the missing PO.

The methodology [155, 160, 154] for estimating causal effects can be broadly divided into two settings: *randomized experiments* and *observational studies*.

21.6.1 Randomized Experiments

In a randomized experiment, treatment assignment T_i is independent of potential outcomes $(Y_i(1), Y_i(0))$. This design guarantees the condition of ignorability by construction:

$$(Y(1), Y(0)) \perp\!\!\!\perp T. \tag{21.15}$$

Under this setting, unbiased estimation of the average treatment effect (ATE) can be achieved directly through sample means:

$$\hat{\tau}_{ATE} = \bar{Y}_{T=1} - \bar{Y}_{T=0}, \tag{21.16}$$

where $\bar{Y}_{T=1}$ and $\bar{Y}_{T=0}$ are the sample averages of observed outcomes in the treated and control groups, respectively. The randomization ensures that differences in sample means consistently estimate $E[Y(1) - Y(0)]$ without further adjustment. The sampling variance can be estimated using standard methods for independent samples, allowing for inference through confidence intervals and hypothesis testing.

21.6.2 Observational Studies

When randomization is not possible, treatment assignment may depend on pre-treatment covariates X. In this setting, unbiased estimation requires conditioning on X to recover the independence between treatment and potential outcomes:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X. \tag{21.17}$$

The estimation then proceeds by comparing treated and control units with identical or similar covariate values.

Several classical estimation strategies fall under this approach:

(a) Covariate Adjustment (Regression Estimation)

A parametric model is specified for the conditional expectation of Y given T and X:

$$\mathbb{E}[Y \mid T, X] = \alpha + \tau T + f(X), \tag{21.18}$$

where $f(X)$ captures the effect of covariates. The coefficient τ provides an estimate of the causal effect, assuming the model is correctly specified and ignorability holds.

(b) Matching Estimation

Units in the treatment and control groups are paired based on similarity in X, thereby approximating the counterfactual outcomes. The causal effect is estimated by averaging outcome differences across matched pairs:

$$\hat{\tau}_{match} = \frac{1}{N_T} \sum_{i \in T=1} (Y_i - Y_{j(i)}), \tag{21.19}$$

where $j(i)$ denotes the matched control unit for treated unit i. Matching directly implements the principle of comparing "statistical twins" under equivalent covariate conditions.

(c) Propensity Score Methods

Rubin and Rosenbaum [151] introduced the *propensity score*—the probability of receiving treatment given covariates:

$$e(X) = P(T = 1 \mid X). \tag{21.20}$$

Conditioning on $e(X)$ rather than X itself preserves the ignorability property:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(X). \tag{21.21}$$

This allows estimation through several equivalent procedures:

- *Stratification or subclassification:* dividing the sample into strata based on estimated propensity scores and computing within-stratum mean differences;
- *Weighting:* using inverse probability weighting (IPW) to balance covariate distributions between groups;
- *Matching:* pairing treated and control units based on the estimated propensity score.

Each approach produces an unbiased estimate of $\mathbb{E}[Y(1) - Y(0)]$ under correct specification of $e(X)$.

21.6.3 Inference and Uncertainty Estimation

Uncertainty in estimated causal effects arises from both sampling variability and the estimation of counterfactuals. Standard errors are computed under the assumption of independent units, and Rubin proposed using repeated sampling logic [156] or the Bayesian posterior variance [155, 158] to quantify uncertainty. In Bayesian implementations, the PO and parameters are treated as random variables, and the posterior distribution of the causal effect provides direct probabilistic statements about uncertainty.

21.6.4 Censoring and Principal Stratification

In many empirical studies, especially in biomedical, social, and economic research, causal inference is complicated by the fact that some post-treatment outcomes may be *undefined* for certain units. This situation, often referred to as *censoring* or *truncation*, occurs whenever the measurement of the outcome of interest is not possible for all units due to intermediate events, which may include death, dropout, non-response, or other absorbing events [159]. Such censoring violates the standard potential outcome assumption that both potential outcomes, $Y_i(1)$ and $Y_i(0)$, are well-defined for every unit i.

To address this problem, Rubin [159] introduced the framework of *principal stratification*, which defines causal effects within subpopulations (principal strata) characterized by the joint potential outcomes of the intermediate post-treatment variable. Let $S_i(1)$ and $S_i(0)$ denote the potential outcome of an intermediate variable under treatment and control, respectively. The pair $(S_i(1), S_i(0))$ defines principal strata that classify

units according to how the intermediate event would manifest under both treatment conditions.

For example, if the intermediate variable is survival, the strata are:

$$\begin{aligned} &\text{Always-survivors:} && S_i(1) = 1,\, S_i(0) = 1, \\ &\text{Protected:} && S_i(1) = 1,\, S_i(0) = 0, \\ &\text{Harmed:} && S_i(1) = 0,\, S_i(0) = 1, \\ &\text{Never-survivors:} && S_i(1) = 0,\, S_i(0) = 0. \end{aligned} \tag{21.22}$$

In general, causal effects on the post-treatment outcome Y are only well-defined for units in strata where $Y_i(1)$ and $Y_i(0)$ are both defined. For the survival example, this corresponds to the *always-survivors* stratum, and the stratum-specific causal effect is:

$$\tau_{AS} = \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) = S_i(0) = 1]. \tag{21.23}$$

Since stratum membership $(S_i(1), S_i(0))$ is never fully observed (each unit experiences only one treatment level), estimation requires assumptions or auxiliary information. Covariates predictive of stratum membership can be used in modeling approaches to infer causal effects probabilistically, for example via Bayesian or likelihood-based methods.

Key considerations in this framework include:

- *Stratum-specific estimands:* When post-treatment outcomes are undefined for part of the population, the overall average treatment effect (ATE) may not be well-defined. Researchers should specify the principal stratum of interest.
- *Separating effects:* Treatments may affect both the intermediate variable (e.g., survival or censoring event) and the post-treatment outcome. Principal stratification conceptually disentangles these effects, though identification often requires strong assumptions.

In practice, inference often involves joint modeling of the intermediate variable and the outcome, or sensitivity analyses to assess the robustness of conclusions under alternative specifications of unobserved strata.

In summary, censoring due to any intermediate post-treatment event poses challenges for causal inference under the PO framework. Principal stratification provides a coherent approach to define and estimate causal effects within well-defined subpopulations, preserving the logic of PO while acknowledging the limits of identifiability in the presence of post-treatment truncation.

21.6.5 Summary

In summary, the PO methodology defines causal effects as contrasts between well-defined potential outcomes and provides a coherent framework for estimating these effects under both randomized and non-randomized designs. Through conditioning, matching, or weighting, it reconstructs the missing counterfactuals under explicit assumptions, yielding interpretable and statistically principled estimates of causal effects. Its strength

lies in its conceptual clarity: it separates the *definition of causality* from the *estimation procedure*, ensuring that all empirical analyses remain grounded in a transparent counterfactual logic.

By framing causal inference as a missing data problem governed by clear assumptions, the PO framework provides a unifying statistical paradigm. RCTs naturally satisfy the ignorability assumption by design, while observational studies require careful adjustment strategies—such as matching, regression, propensity scores, or instrumental variables, which are discussed in Chapter 22—to approximate randomization. This versatility explains why the PO framework has become the dominant language of modern causal inference, offering both clarity in conceptual reasoning and rigor in statistical estimation.

21.7 Example: Infer the Effect of the Apple Origins on the Purchasing Prices

This example reformulates the apple purchasing price problem under the PO framework originally proposed by Neyman (1923) and Rubin (1974). The goal is to estimate the causal effect of *apple origin (PL)* on the *purchasing price (APP)* of apples, both directly and indirectly through key mediating characteristics: *sweetness (SW)*, *size (AS)*, and *shelf life (SL)*.

Units, Treatment, and Potential Outcomes

Each apple batch or shipment represents a distinct *unit* $i \in \{1, 2, \ldots, N\}$. The *treatment variable* PL_i denotes the *origin status* of apples, representing whether they come from a favorable cultivation region (e.g., advantageous climate or soil). For simplicity, we assume two treatment levels:

$$PL_i \in \{0, 1\}, \quad 0 = \text{less favorable origin}, \quad 1 = \text{favorable origin}.$$

For each unit i, the PO for purchasing price is defined as:

$$Y_i(PL_i) = \textit{Purchasing price that would be observed if treatment } = PL_i.$$

However, only one of these potential outcomes is observed:

$$Y_i^{obs} = Y_i(PL_i^{obs}),$$

while the counterfactual outcome $Y_i(1 - PL_i^{obs})$ remains unobserved.

Mediators and Outcome

To capture the mechanisms through which *the apple origin* influences the purchasing price outcome, we define a vector of mediating variables for each unit:

$$M_i = (SW_i, AS_i, SL_i),$$

corresponding to sweetness, size, and shelf life.
Each mediator is itself affected by the treatment, and thus has its own potential outcomes:

$$SW_i(PL_i), \quad AS_i(PL_i), \quad SL_i(PL_i).$$

The PO for purchasing price can therefore be represented as:

$$Y_i(PL_i, M_i(PL_i)),$$

indicating that the apple origins may influence purchasing price both directly (via PL_i) and indirectly (via M_i).

Causal Estimands

Within the PO framework, the *Average Treatment Effect (ATE)* of the apple origin on the purchasing price is defined as:

$$ATE = \mathbb{E}[Y_i(1) - Y_i(0)],$$

which represents the expected causal effect on the purchasing price when switching from a less favorable to a favorable origin.
To decompose this total effect into direct and indirect components, we define the *Natural Direct Effect (NDE)* and *Natural Indirect Effect (NIE)* as:

$$NDE = \mathbb{E}[Y_i(1, M_i(0)) - Y_i(0, M_i(0))],$$

$$NIE = \mathbb{E}[Y_i(1, M_i(1)) - Y_i(1, M_i(0))].$$

- The *NDE* measures how the price would change if the treatment changed, but mediators were fixed at their counterfactual values under the less favorable origin. - The *NIE* measures how the price changes solely due to treatment-induced changes in mediators, holding the treatment level fixed.
By definition:

$$ATE = NDE + NIE.$$

Covariates and Assignment Mechanism

We denote X_i as a vector of pre-treatment covariates (e.g., production scale, harvest period, market conditions) that may confound the treatment–outcome relationship. The *assignment mechanism* $P(PL_i = 1 \mid X_i)$ governs how treatment is assigned given these covariates. Proper adjustment for X_i ensures that the causal estimands are identifiable under standard assumptions.

Assumptions for Identification

Identification of ATE, NDE, and NIE requires the following PO assumptions:

- *Stable Unit Treatment Value Assumption (SUTVA):* Each unit's PO depends only on its own treatment and mediators, implying no interference between units and no multiple treatment versions:

$$Y_i(PL_i, M_i(PL_i)) \text{ and } M_i(PL_i) \text{ are well-defined for all } i.$$

- *Ignorability (Unconfoundedness):* Conditional on covariates X_i, treatment assignment is independent of PO and mediators:

$$\{Y_i(1), Y_i(0), M_i(1), M_i(0)\} \perp\!\!\!\perp PL_i \mid X_i.$$

- *Positivity:* Each treatment level occurs with positive probability given the covariates:

$$0 < P(PL_i = 1 \mid X_i) < 1.$$

Numerical Illustration

Assume a binary treatment scenario with hypothetical means calibrated from agricultural data:

$$\mathbb{E}[Y_i(1)] = 6.8 \text{ ¥/kg}, \quad \mathbb{E}[Y_i(0)] = 6.1 \text{ ¥/kg}.$$

Then:

$$ATE = 6.8 - 6.1 = 0.7 \text{ ¥/kg}.$$

Further decomposition yields:

$$NDE = 0.25 \text{ ¥/kg}, \quad NIE = 0.45 \text{ ¥/kg}.$$

Thus, about 64% of the total effect is mediated through improvements in *sweetness, size, and shelf life*, while 36% reflects the direct effect attributable to origin-specific market advantages.

Interpretation

The PO framework explicitly distinguishes *potential outcomes* from *observed outcomes*, enabling formal reasoning about *counterfactuals*—for example, the purchasing price an apple batch would have had if grown in a different origin.
This causal decomposition clarifies that:

- The *direct effect* captures regional or brand-related advantages associated with favorable origins;
- The *indirect effect* captures quality improvements—greater sweetness, larger size, and longer shelf life—induced by better cultivation conditions.

Overall, even under identical market conditions, transitioning to a favorable origin raises average apple prices by approximately 0.7 ¥/kg, primarily through mediator-driven quality effects.

21.8 Limitations

The PO framework provides a rigorous foundation for defining and estimating causal effects. However, several limitations should be acknowledged.

First, the framework relies on key identification assumptions, such as *ignorability* (no unmeasured confounding), *positivity*, and the *Stable Unit Treatment Value Assumption (SUTVA)*, which are often untestable in practice. Violations of these assumptions, for instance, due to hidden confounders or interference between units, can lead to biased causal estimates.

Second, PO-based analyses are typically data-intensive: reliable estimation of counterfactual quantities requires large and well-balanced samples, particularly when the treatment or mediator is continuous or high-dimensional.

Third, while the framework conceptually separates treatment, outcome, and covariates, it cannot on its own determine the correct causal structure; misspecifying covariates or mediators may distort the estimated causal effects. Furthermore, estimation of indirect (mediated) effects introduces additional assumptions, such as sequential ignorability, that are even harder to verify empirically.

Finally, although the PO framework offers a precise language for causal reasoning, it does not itself provide causal identification without strong assumptions or experimental control. Therefore, causal interpretations under the PO framework should always be supported by substantive theory, domain knowledge, and sensitivity analyses.

21.9 Summary

The Potential Outcome framework is a fundamental approach for defining and estimating causal effects. It formalizes causal questions by comparing outcomes that would be observed under different treatment conditions, thereby reframing causal inference as

a missing-data problem. The Potential Outcome approach is especially powerful because it connects causal reasoning to explicit assumptions, such as consistency, SUTVA, and ignorability, that are required to recover unobserved counterfactuals from available data. When applied carefully, it enables researchers to quantify causal effects, evaluate treatment strategies, and understand heterogeneity across populations. Nevertheless, rigorous causal inference within this framework depends critically on study design, assumption validity, and the quality of observed covariates. Violations such as unmeasured confounding, interference, or selection bias can compromise the interpretation of estimated causal effects. Thus, while the Potential Outcome framework provides a robust theoretical foundation for causal analysis, its practical success requires disciplined implementation and careful attention to methodological detail.

Chapter 22

Instrumental Variables

This chapter [1] presents the basic concepts, problem statement, assumptions, principles, methodology, and case study of the instrumental variables (in short, IV).

22.1 Basic Concepts

An Outcome Variable (Y) is the dependent variable whose causal relationship with the endogenous variable X under investigation [217].

An Endogenous Variable (X) is the independent variable presumed to have a causal impact on an outcome variable Y [217].

Ordinary Regression is a statistical method that fits a linear relationship between the explanatory variable and the outcome variable by minimizing the sum of squared differences between observed and predicted values of the outcome variable [53]. Ordinary regression is typically estimated using ordinary least squares. In IV, the endogenous variable X is the explanatory variable, and Y is the outcome variable.

Endogeneity refers to the situation where an explanatory variable in a statistical model is correlated with the error term [217]. For instance, endogenous explanatory variable X may be correlated with unobserved confounders, causing endogeneity [141].

An IV (Z) serves to isolate the component of variation in the endogenous variable X that is exogenous to the outcome variable Y, and the IV must also be independent of any unobserved confounders [141].

22.2 Problem Statement

IV is a statistical method commonly used in econometrics and data analysis to estimate causal relationships, particularly when conducting controlled experiments is not feasible [217, 149, 166, 141].

The IV problem could be stated as "how to address the *endogeneity* in estimating the causal influence of variable X on variable Y, formally expressed as $Y = \alpha + \beta X + \epsilon$."

[1] The primary contributor is Dr. Lei Wang.

Endogeneity arises when X is correlated with the error term ϵ, leading to biased and inconsistent estimates of β.

The input to the IV problem consists of observed data on the endogenous variable X, the outcome Y, and a proposed IV Z. A valid solution critically depends on whether Z satisfies three key constraints, detailed in Section 22.3. When these constraints are met, the output of the IV procedure is a consistent estimate of the causal parameter β.

22.3 Basic Assumptions

Assumption 22.1 (Relevance)**.** *An IV Z must be correlated with the endogenous variable X. That is, the IV must have a non-zero impact on X (the IV must shift X significantly).*

Assumption 22.2 (Exogeneity)**.** *An IVZ must be uncorrelated with the error term. This ensures that the variation in X induced by Z is exogenous and does not suffer from the same bias as the endogenous variable X.*

Assumption 22.3 (Exclusion Restriction)**.** *An IV Z must only influence the outcome Y through its influence on X and should have no direct influence on Y. If Z directly affects Y, then the estimate of the causal influence will be biased.*

22.4 Basic Principles

IV provides a powerful solution to the problem of *endogeneity* in causal inference. Endogeneity arises when an explanatory variable X is correlated with the error term in a regression model, leading to biased estimates of its causal effect on an outcome Y. A primary source of endogeneity is *confounding*—when unobserved variables U influence both X and Y, creating a spurious association. The IV method addresses this by introducing an instrument Z that serves as an exogenous source of variation. As shown in Figure 22.1, the core idea is to use a IV Z that is correlated with the endogenous variable X but independent of the confounders U, and that affects the outcome Y only through X.

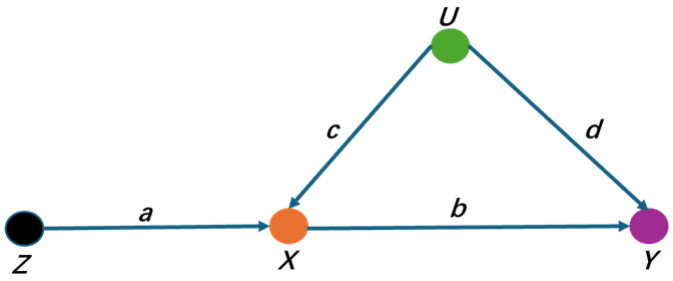


Figure 22.1: The IV framework: An IV Z addresses endogeneity by providing exogenous variation in X. Z is independent of confounders U [141].

The most common implementation of IV is *two-stage least squares*. This procedure involves two stages:

- *First stage*: The endogenous variable X is regressed on the IV Z (and any other exogenous controls) to obtain its predicted values, $\hat{X}$.
- *Second stage*: The outcome Y is regressed on the predicted values $\hat{X}$. This step isolates the component of X that is driven by the exogenous IV, thereby purging the bias caused by confounders U.

While confounding is a major application, IV methods are also essential for addressing other forms of endogeneity, including:

- *Measurement error*: When X is measured with error, leading to attenuated estimates.
- *Simultaneity (reverse causality)*: When X and Y mutually influence each other.
- *Omitted variables*: A general case encompassing confounding, when any unobserved factor affects both X and Y.

Historical applications—from John Snow's investigation of cholera transmission to Philip Wright's analysis of supply curves—demonstrate how well-chosen IV can uncover causal relationships [141]. In summary, the IV framework provides a rigorous approach for causal estimation in non-experimental settings. However, its validity critically depends on satisfying the core assumptions of IV relevance, exogeneity, and the exclusion restriction. Violations of these assumptions can lead to misleading conclusions, necessitating careful design and testing.

22.5 Methodology

The basic model of IV is a regression model. Suppose we have a regression model where Y is the dependent variable, X is the endogenous explanatory variable, and Z is IV. The basic model we wish to estimate is:

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{22.1}$$

- Y: The outcome variable.
- X: The endogenous explanatory variable.
- β_0: Intercept term, representing the expected value of Y when $X = 0$.
- β_1: The coefficient of X, representing the impact of a one-unit change in X on Y.
- ϵ: Error term, representing all factors or random errors not explained by X.

However, due to endogeneity (i.e., X is correlated with the error term ϵ), we cannot directly estimate β_1. To address this issue, we introduce an IV Z, which is correlated with X but uncorrelated with ϵ. The standard estimation procedure of IV is *two-stage*.

Stage 1: Regressing the Endogenous Explanatory Variable In the first stage, we predict X using Z, i.e., the regression:

$$X = \pi_0 + \pi_1 Z + u. \tag{22.2}$$

- X: The endogenous explanatory variable.
- Z: IV.
- π_0: Intercept term, representing the expected value of X when $Z = 0$.
- π_1: The coefficient of IV Z, representing the impact of a one-unit change in Z on X.
- u: Error term, representing the part of X not explained by Z.

The goal of this regression is to explain the variation in X through IV Z.

Stage 2: Using the Predicted Values for Regression In the second stage, we replace X with the predicted value $\hat{X}$ obtained from the first stage regression (which is a linear combination of Z), and then estimate the relationship between Y and $\hat{X}$:

$$Y = \beta_0 + \beta_1 \hat{X} + \nu. \tag{22.3}$$

- Y: The outcome variable.
- $\hat{X}$: The predicted value of X obtained from the first-stage regression, which is a linear combination of Z.
- β_0: Intercept term, representing the expected value of Y when $\hat{X} = 0$.
- β_1: New regression coefficient, representing the impact of the predicted $\hat{X}$ on Y in the second stage regression.
- ν: New error term, representing the random error component in the second-stage regression, which is uncorrelated with ϵ.

Since $\hat{X}$ is predicted using Z and Z is not correlated with ϵ, this second stage regression provides a consistent estimate of β_1.

22.6 Example: Infer the Effect of the Apple Origins on the Purchasing Prices

This example illustrates the application of the IV method to address endogeneity arising from an *unobserved confounder*. We aim to estimate the causal effect of *the apple origin* on the purchase price. The primary source of endogeneity is *unobserved consumer perception* (e.g., the belief that local apples are fresher or higher quality). This perception influences both the likelihood of an apple being marketed as local and the price consumers are willing to pay. Since we cannot measure this perception directly, we use an IV to isolate exogenous variation in origin.

Variables

- Y: purchase price (per kilogram, in ¥).
- X: Apple origin (binary variable, 0 = other regions, 1 = local origin). This is the *endogenous* variable.
- Z: Agricultural policy for subsidized farming (binary variable, 0 = no policy, 1 = policy in place). This is an IV.
- U: *Unobserved* consumer perception of local produce quality.

Hypothetical Data

We collected data from 1,000 apple shipments.

- Y: Apple purchase price ranges from 5.00 to 15.00 ¥/kg.
- X: Apple origin is 0 (Other regions) or 1 (Local).
- Z: Agricultural policy is 0 (No policy) or 1 (Policy in place).

Assume the following sample data: - 500 samples from the apples from the other regions ($X = 0$) - 500 samples from the locally grown apples ($X = 1$) - 60% of the samples (600 shipments) are from the regions subject to the agricultural policy ($Z = 1$)

Problem Setup: Unobserved Confounding

We wish to estimate the causal model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where the error term ϵ contains the unobserved consumer perception U. This creates endogeneity because U influences both X (Apple origin) and Y (Purchasing price), leading to a correlation between X and ϵ. For example, regions with strong positive perception (U) will have more local apples (X) and higher prices (Y), creating a

spurious correlation. An ordinary least squares regression of Y on X would yield a biased estimate of β_1.

An IV Solution

We introduce an IV Z, an agricultural subsidy policy (the government program that reduces the cost of local agricultural production to encourage local farming), to resolve the endogeneity. The validity of Z rests on two assumptions:

- *Relevance:* The policy Z directly influences the apple origin X. Subsidies lower the cost of local production, making it more likely for apples to be grown and sold locally ($X = 1$). Hence, Z is correlated with X.
- *Exclusion Restriction:* The policy Z affects the purchasing price Y *only* through its effect on the supply of local apples X. The policy itself is unrelated to unobserved consumer perception U, and does not directly affect market prices other than by shifting the origin composition.

IV provides a source of exogenous variation in X that is independent of the confounder U.

Two-Stage Least Squares Implementation

Stage 1: Isolating Exogenous Variation in the Apple Origins
In the first stage, we regress the endogenous variable X on the IV Z. This isolates the variation in the apple origin that is driven solely by the exogenous policy:

$$X_i = \pi_0 + \pi_1 Z_i + u_i,$$

where X_i indicates the local origin (1) or not (0), and Z_i indicates the presence of the policy (1) or not (0). The regression is estimated using ordinary least squares, producing the fitted values $\hat{X}_i$, which represent the component of the apple origin that is driven by the policy and is therefore exogenous to the unobserved confounders in ϵ.
Stage 2: Estimating the Causal Effect on Purchasing Price
In the second stage of the two-stage least squares procedure, we regress the purchasing price Y_i on the predicted (exogenous) values of the apple origin $\hat{X}_i$ from the first stage:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + \nu_i.$$

This stage yields a consistent estimate of the causal parameter β_1, as it uses only the exogenous variation in X induced by the IV Z.
Illustrative Numerical Results
Assume the first-stage ordinary least squares regression yields:

$$X_i = 0.20 + 0.65 Z_i + u_i$$

This suggests the policy increases the probability of the local apple origin by 65 percentage points.
Substituting $\hat{X}_i$ into the second-stage regression gives the two-stage least squares estimate:

$$Y_i = 8.50 + 2.10\hat{X}_i + \nu_i.$$

This implies that the causal effect of an apple being locally grown is to increase its price by an average of *2.10 ¥/kg*.

Conclusion and Interpretation

The two-stage least squares procedure provides a consistent estimate of the causal impact of the apple origins on the purchase prices. The first stage isolates exogenous variation in the apple origin due to the agricultural policy, while the second stage corrects for endogeneity and yields consistent parameter estimates. Under the assumptions of instrument relevance and exclusion, the coefficient $\hat{\beta}_1 = 2.10$ indicates that *locally grown apples are priced approximately 2.10 ¥ higher per kilogram* than non-local apples.
However, it is important to note that this estimate represents the Local Average Treatment Effect (LATE), which applies specifically to the subset of apples whose origin is influenced by the agricultural policy. This means that the estimated effect reflects the price difference between locally grown and non-local apples only in regions where the policy has altered the origin of apples. In other words, this result is not a general estimate for all local apples but specifically for those whose production decisions are affected by the policy. Thus, the estimate $\hat{\beta}_1 = 2.10$ reflects the causal effect of being locally grown on price for apples in areas where the policy has induced changes in production decisions.

The above example demonstrates how IV can be used to address endogeneity issues and estimate the causal impact of the apple origin on the purchase price, while also verifying the indirect impact of an agricultural policy on purchasing prices. Through this method, we can draw more reliable conclusions about the impact of an agricultural policy on market outcomes, such as apple pricing.

22.7 Limitations

Despite their power, IV methods have important limitations. First, finding a valid IV is difficult because few variables in practice simultaneously satisfy relevance, exclusion, and exogeneity. Second, a weak IV, which is only weakly correlated with the variable of interest (X), can lead to biased estimates and large standard errors. Third, IV estimates typically identify a local average treatment effect (LATE) for the subpopulation affected by IV, limiting the generalizability of the results. Finally, IV estimates are sensitive to model specification, and incorrect models or inappropriate covariates can compromise the validity of the estimates.

22.8 Summary

Instrumental Variables (IV) methods are widely used in econometrics to address endogeneity and estimate causal relationships when controlled experiments are not feasible. By using an instrument—a variable correlated with the endogenous explanatory variable but not with the error term—IV methods help isolate the causal effect of the explanatory variable on the outcome. However, IV methods have limitations, including challenges in finding valid instruments, sensitivity to model specification, and the potential for weak instruments to lead to biased estimates.

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

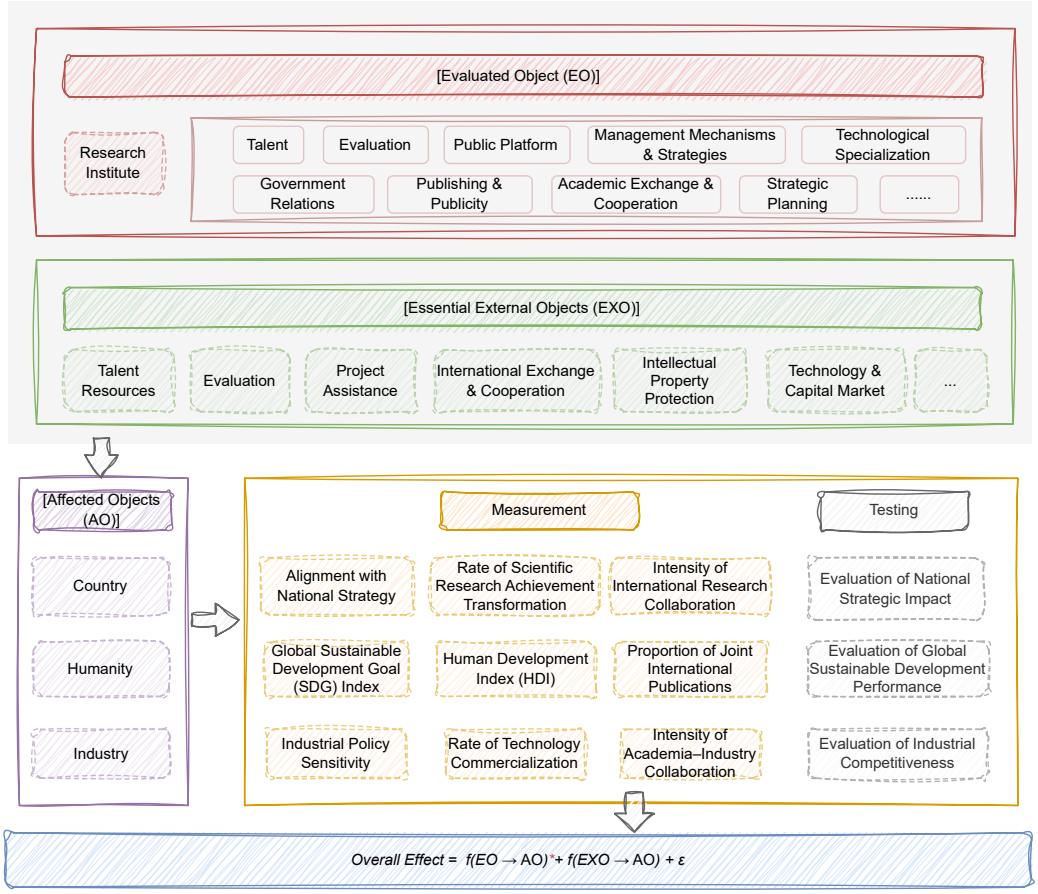


Figure 23.1: An SES for Evaluating an STRI.

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 $\widehat{Y}_{\mathrm{AO}}$ denote the measured effect on an AO, which aggregates contributions from multiple pathways:

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

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 $\widehat{Y}_{\mathrm{AO}}$, 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:

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

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.

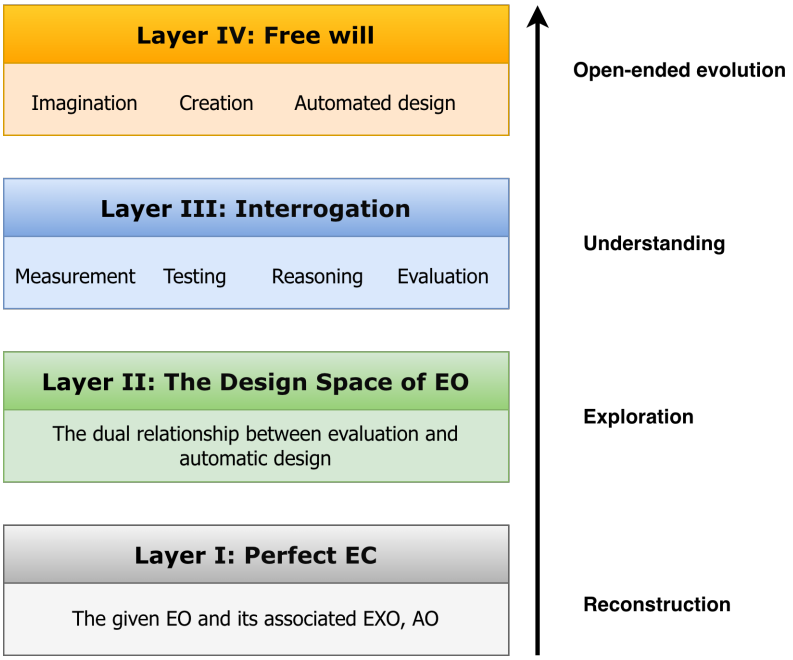


Figure 25.1: Evaluatology-based pathway toward strong AI.

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 designed. For example, it can refer to the model or algorithm itself—such as a video retrieval network, an encoder-decoder architecture, or a recommendation algorithm. In this context, its internal design factors, including model architecture and learning strategy, determine the nature and magnitude of its influence. The EO could be a CPU or a database system.

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* represents the measurable outcome or behavior influenced by both the EO and EXO. 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 EXO.

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, EXO, and AO.

Step I: Design and Implement a Perfect EC: At this foundational level, for any given EO and its associated EXO 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 EXO 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 EXO, 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 deliberate optimization and selection of designs that best fulfill stakeholders' evaluation requirements. 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 EXO. 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 *EXO* consists of all factors that influence index performance. The EXO includes, but is not limited to, data distribution and skew patterns, schema evolution and update frequencies, storage layout, and compression rules. Inspired by CPU Evaluatology, the EXO is 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] Emest 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.