评价学: 评价科学和工程
Evaluatology: The Science and Engineering of Evaluation

詹剑锋[a,b,c,1], 王磊[b,a,c] and 高婉铃[b,a,c]

[a] 国际测试委员会
[b] 中国科学院计算技术研究所
[c] 中国科学院大学

ARTICLE INFO

Keywords:
评价（Evaluation）
评价基准 (Benchmark)
量表（Scale）
指数（Index）
评价条件 (Evaluation Condition)
评价模型 (Evaluation Model)
评价系统 (Evaluation System)
评价标准 (Evaluation Standard)
等价评价条件 (Equivalent Evaluation Condition)
最小等价评价条件 (Least Equivalent Evaluation Condition
评价学 (Evaluatology)
基准学 (Benchmarkology)

摘要

评价是人类基本活动之一，在各个领域中扮演着至关重要的角色。然而，不同领域的评价通常是经验性的，与具体场景相关 (ad-hoc)，并且缺乏统一的评价概念、术语、理论和方法。这种共识的缺失可能会导致严重的后果。本文旨在正式介绍评价学 (Evaluatology) 这一学科，它涵盖了评价科学和评价工程。评价科学的核心问题是："任何评价结果是否具有真值?"评价工程则关注如何在满足利益相关者 (stakeholders) 的评价需求的同时，最小化评价成本。为了解决上述挑战，我们提出了一个通用的评价框架，包括统一的评价学概念 (Concepts)、术语 (Terminologies)、理论 (Theories) 和方法 (Methodologies)。这个通用的评价框架旨在适用于不同学科的评估问题，即便不适用于所有学科。

本文是对评价学的简要总结，若要更全面地了解，请参阅原文https://www.sciencedirect.com/science/article/pii/S2772485924000140。如果您希望引用此工作，请引用原文。
Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, Hainan Ye, Shaopeng Dai, Zhifei Zhang (2024). Evaluatology: The science and engineering of evaluation. BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 4(1), 100162.

1. 引言

评价是人类基本活动之一，在各个领域中扮演着至关重要的角色。然而，不同领域的评价通常是经验性的，与具体场景相关 (ad-hoc)，并且缺乏统一的评价概念、术语、理论和方法。这种共识的缺失可能会导致严重的后果。即使在成熟的计算机工程领域，同一对象（Subject，评价对象）的评价有着显著差异的评价结果 (Evaluation outcome) 并不罕见，这些差异可能导致相互矛盾的结论。例如，使用多个公认的处理器评价基准 (Benchmark) 评价同一处理器的性能时，其评价结果存在巨大差异，而且使用不同评价基准产生的评价结果之间相互无法比较 (comparable)。因此，每个评价人员自然会提出一个评价的核心问题："任何一次评价的结果是否具有真值 (True value)"？尤其涉及安全攸关、任务攸关或者商业攸关场景下的评价时，评价的可靠性、有效性和效率尤为重要，且引人担忧。

本文首次正式介绍了评价学这一学科，它涵盖了评价科学和工程。评价科学的核心问题是："任何评价结果是否具有真值?"评价工程旨在解决如何在满足利益相关者评价需求的同时，最小化评价成本。为了解决上述挑战，我们提出了一个通用的评价框架，包括跨学科适用（即便可能不是所有学科）的概念、术语、理论和方法。图 1展示了评价学中的通用概念、理论和方法。

2. 评价科学 (The science of evaluation)

2.1. 评价的本质 (The essence of evaluation)

评价的挑战源于一个固有事实，即孤立地评估一个对象无法满足利益相关者的评价需求（Evaluation requirements)。相反，通过施加一个明确定义的等价条件（EC）来反映利益相关者的关注或利益至关重要。因此，评价的本质在于给待评价的个体或系统（我们称为评价对象，Subject）施加一个明确定义的评价

jianfengzhan.benchcouncil@gmail.com（詹剑锋）
www.zhanjianfeng.org（詹剑锋）
ORCID(s):

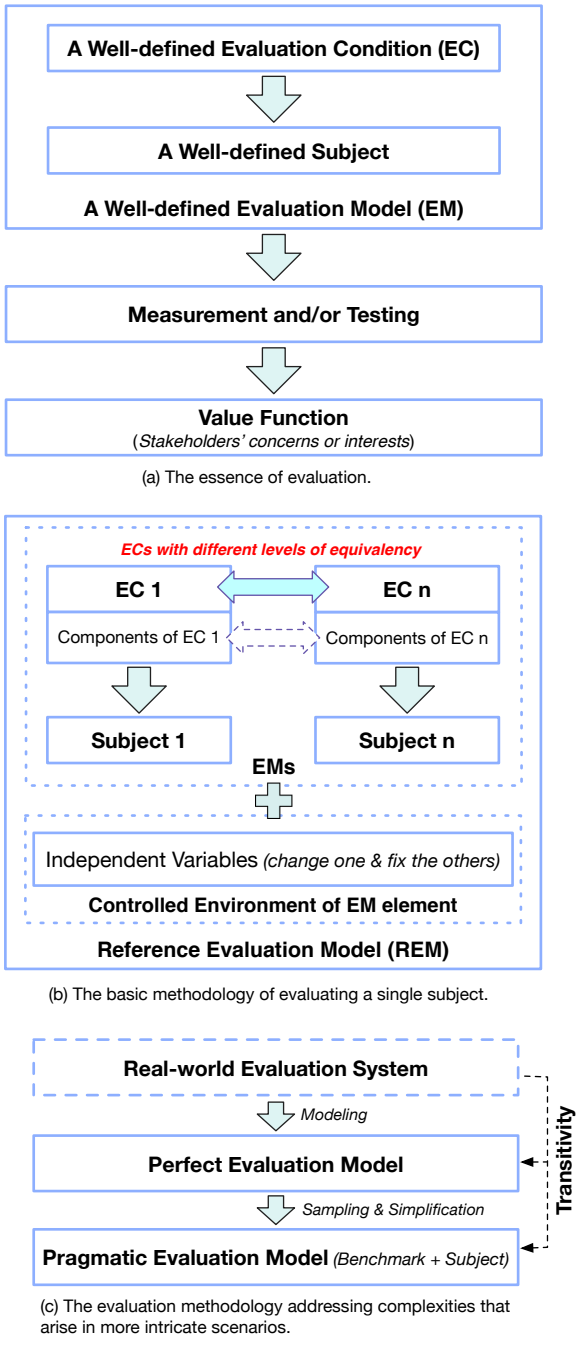

图 1: 评价学的通用概念、理论和方法。

条件（evaluation condition，简称 EC），并进行实验的过程。这一过程即构建了一个评价系统（evaluation system，简称 ES）或者评价模型（evaluation model，简称 EM）。通过对这个评价系统或模型进行计量（measurement）和/或测试（testing），我们间接推断出不同对象的影响，从而对对象进行评价。

2.2. 五大评价公理 (evaluation axioms)

基于评价的本质，我们提出了五大公理作为基础评价理论，这些公理关注评价结果的重要属性，构成了我们构建通用评价理论和方法的基石。

综合评价指标本质公理 (The Axiom of the Essence of Composite Evaluation Metrics)：综合评价指标的本质要么具有内在的物理意义，要么完全由价值函数 (value function) 决定。

评价结果真值公理 (The Axiom of True Evaluation Outcomes)：当一个明确定义的评价条件 (Evaluation condition, EC) 施加于一个明确定义的对象时，其评价结果，包括量和综合评价指标，具有真值。

评价可溯源性公理 (The Axiom of Evaluation Traceability)：对于同一对象，评价结果的差异可归因于评价条件的差异，从而建立评价的可溯源性。

评价结果可比较性公理 (The Axiom of Comparable Evaluation Outcomes)：当每个明确定义的对象都被施加等价的评价条件 (Equivalent evaluation condition, ECC) 时，其评价结果是可比较的。

评价结果一致性公理 (The Axiom of Consistent Evaluation Outcomes)：当对一个明确定义的对象施加评价条件样本 (samples) 时，其评价结果一致地趋近于施加评价条件总体 (Population) 获得的评价结果真值 (True quantity)。

2.3. 基础评价理论 (Basic evaluation theory)

基于上述五个评价学公理，我们提出了通用评价理论和方法。

2.3.1. 评价条件的层次化定义 (The hierarchical definition of an EC)

在进行有意义的评价之前，建立一个明确定义的评价条件是必要的前提。我们提出了一个通用和层次化的评价条件定义方法，并从上到下确定了评价条件的五个核心组件。我们从利益相关者面临和需要解决的问题 (Problem) 或任务 (Task) 空间开始定义评价条件，原因有二。首先，利益相关者的关注和兴趣是评价的核心，这些关注和兴趣最好通过他们必须面对和解决的问题或任务来反映，这提供了定义评价条件的可靠手段。其次，利用相同的问题或任务可以确保评价结果的可比性。

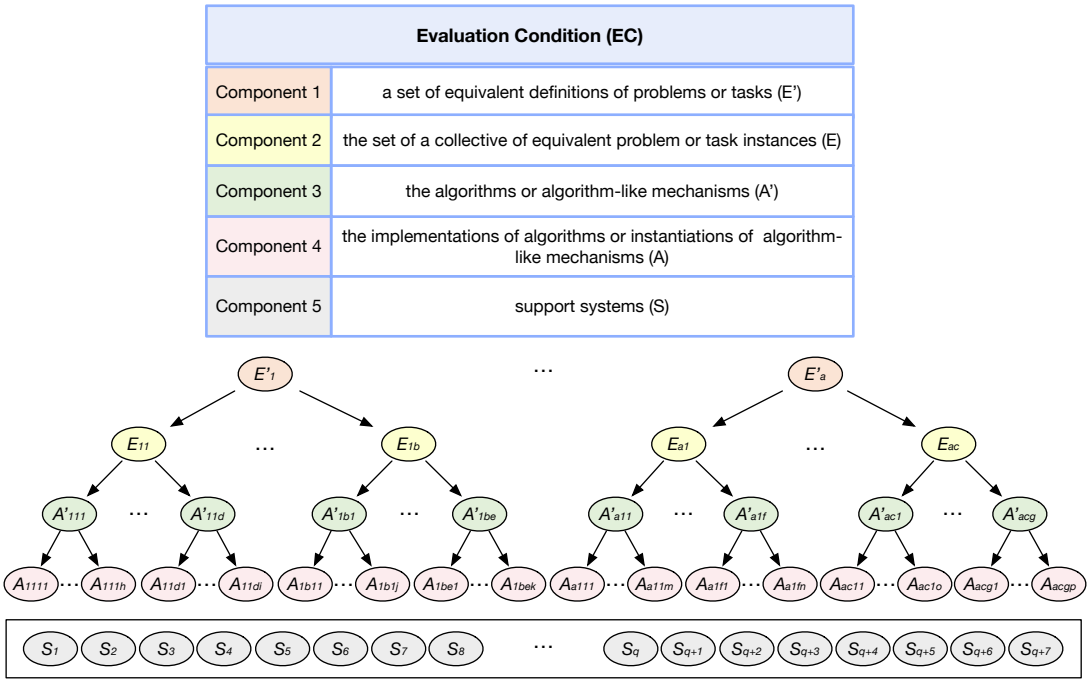

图 2: 评价条件的层次化定义。

虽然问题或任务本身是评价的基础，但它不能单独构成评价，因为问题或任务通常是抽象的，需要进一步实例化以确定其具体参数。第二个核心组件是一组等价的问题或任务实例 (Problem or task instance)，每个实例都是从第一个核心组件的元素实例化而来。不同于第一个核心组件，等价的问题或任务实例是具体的，可以直接用于评价。提出问题或任务实例后，有必要找出解决方案。第三个核心组件包括算法 (Algorithm) 或类算法机制 (Algorithm-like mechanism)，作为特定问题或任务实例的解决方案，其中，类算法机制是指以类似于算法的方式运行的过程。第四个核心组件包括算法的实现或类算法机制的实例化。第五个核心组件是支撑系统，它提供了评价所依赖的必要资源和环境。

2.3.2. 建立等价评价条件或最小等价评价条件 *(The establishment of EECs or LEECs)*

在评价不同对象时，优先使用等价评价条件(EEC)是最重要的。等价评价条件意味着两个评价条件中每一层核心组件都必须是等价的。通过在每一层保持等价性，我们可以确保公平和无偏颇的评价，从而在不同对象之间进行有意义的比较和评价。在某些情况下，实现两个评价条件在所有组件层次上的完全等价具有挑战性，甚至无法实现。面对这一挑战，我们提出了最低级别的等价评价条件，即确保两个评价条件中最重要的组件的等价性，我们称之为最小等价评价条件 (Least equivalent evaluation condition, LEEC)。我们在评价条件的第一和第二个顶层组件的基础上建立最小等价评价条件。

为了建立最小等价评价条件，我们需要确定评价条件 (五个层次的组件) 中必须保证等价性的最重要的一个组件，这个最重要的组件称为评价标准 (Evaluation standard)，它在定义最小等价评价条件中起着至关重要的作用。一个评价标准应该具有三个最基本的属性：可求解 (Solvable)、明确定义 (Definite) 和等价性 (Equivalent)。根据英文的首字母缩写，将这三个属性简称为 SDE 属性。我们将评价标准定义在评价条件的第二个顶层组件上：即一个具体的问题或者任务实例。

2.3.3. 建立参考评价模型 *(The establishment of an REM)*

在构建评价模型 (EM) 时，我们在不同评价对象上施加不同等价性的评价条件。评价模型中的一个元素是评价模型状态空间中的一个特定点，每个元素可能有许多独立变量 (Independent variables)。为了消除混杂因素的影响 (Confounding)，我们提出了一个称为参考评价模型 (REM) 的新概念。参考评价模型

要求每个元素在变化时一次只改变一个独立变量，同时将其他独立变量作为控制变量 (Control variable)。随后，我们利用计量和/或测试来评价参考评价模型的功能。最后，通过搜集和分析评价系统的计量和测试数据，我们推断出不同对象的因果（cause–effect）影响。

2.4. 复杂场景下的通用评价方法学 (Universal evaluation methodology in complex scenarios)

应对复杂场景下的评价挑战，我们揭示了有效且高效的评价的关键在于建立一系列保持传递性(transitivity）的评价模型。在完整版本里 [15]，我们用数学工具形式化地定义了传递性。

我们将用于评价特定对象的真实世界系统总体(population) 称为真实世界评价系统(real-world ES)。假设不存在安全问题，真实世界评价系统是构建最佳评价环境的首要候选，适用于不同对象的评价。然而，使用真实世界评价系统具有一系列重大障碍，例如大量混杂因素 (confounding)、建立参考评价模型的挑战、巨大的评价成本、大量无关的并发问题或任务、以及状态空间内对某些评价条件聚类的偏向性(inclination to exhibit bias towards certain clusters within the EC state space)。

我们假设存在完美评价模型（perfect EM），它以最完美的方式复制真实世界评价系统。完美的评价模型消除无关的问题或任务，能够全面探索和理解评价条件的所有可能性，并促进参考评价模型的建立。然而，完美评价模型具有巨大的状态空间，涉及大量独立变量，因此导致评价成本高昂。为了解决这一问题，我们进一步提出实用评价模型（pragmatic EM）的概念以简化完美评价模型：一方面消除结果影响小的独立变量从而减少独立变量数目；另一方面，通过对巨大的状态空间进行采样 (Sampling)，可以减小空间。实用评价模型提供了一种估计真实世界评价系统参数 (Parameter) 的方法。

2.5. 评价学的基础问题 (Fundamental issues in evaluatology)

我们提出了评价学的四个基本问题，并在完整版本里 [15] 用数学工具形式化地描述了这四个基本的评价学问题：

(1) 如何确保评价模型的传递性(transitivity) 是在复杂场景下构建评价模型的最基本问题之一。这个问题涉及到在真实世界评价系统基础上构建完美评价模型，并进一步简化以得到实用评价模型。

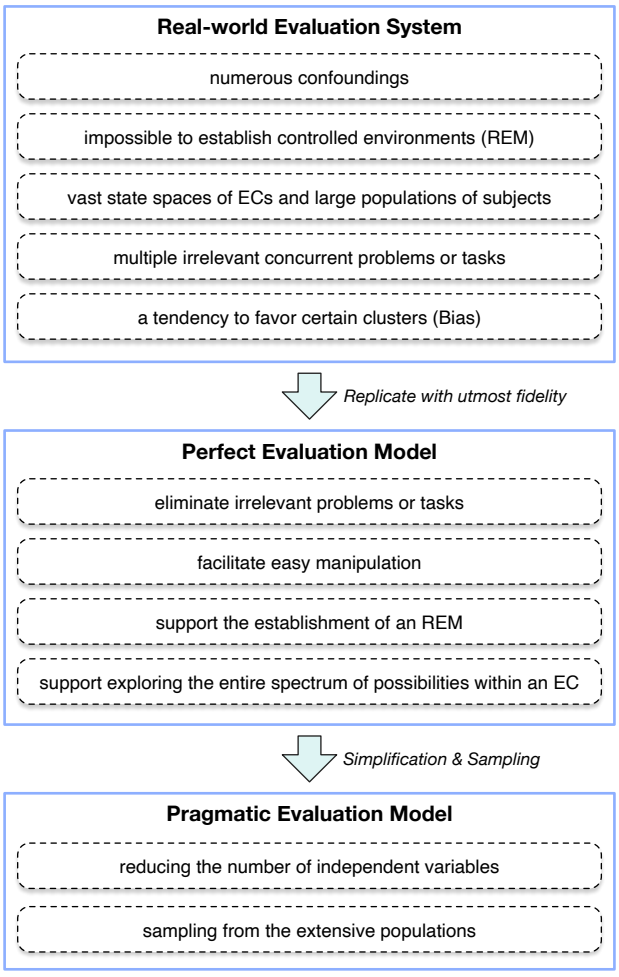

图 3: 复杂场景的通用评价方法。

（2）如何控制评价结果在接近真值的某个范围内（controlled discrepancies）同时降低评价成本是评价过程中最重要的工程问题，即在不超过评价结果差异阈值 (discrepancy threshold of the evaluation outcomes）的条件下最小化评价成本。

（3）确保评价的可溯源性 (Evaluation traceability) 是一个需要应用科学和工程原理的多方面的问题。它涉及将评价结果的任何差异归因于评价条件的差异，从而建立清晰透明的可溯源性。

(4) 如何在各学科的评价标准之间建立连接和关联，是评价学的大统一理论(the grand unified theory of evaluatology)，这使得对评价相关问题进行全面探讨成为可能。评价标准是任何评价模型中的基础支柱。通过建立不同学科之间评价标准的联系，我们

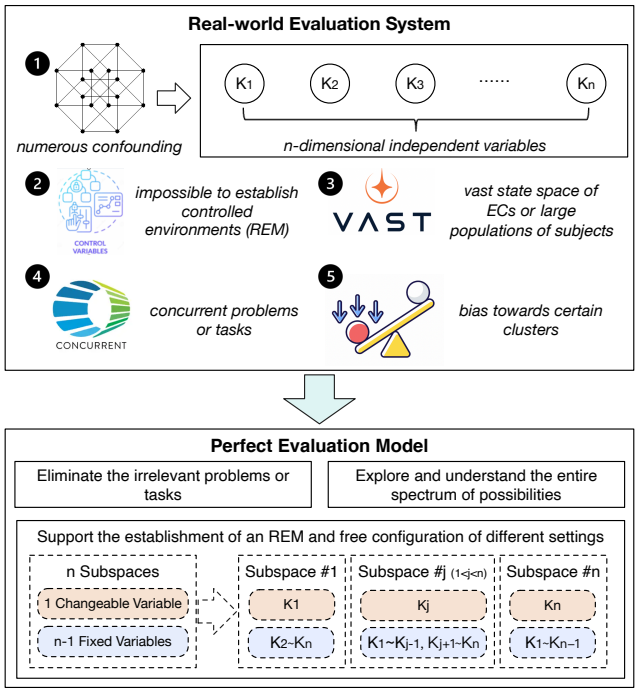

图 4: 完美评价模型复刻真实世界评价系统。

有可能构建一个覆盖所有领域评价问题的综合框架。

3. 基准学：评价工程学 (Benchmarkology: the engineering of evaluation)

评价基准（Benchmark）尽管缺乏正式的定义，仍被广泛应用于很多学科领域。基于评价科学，我们提出了对评价基准的精确界定，将其描述为简化和采样的评价条件（EC），即实用评价条件 (pragmatic EC)，它确保了从最小等价评价条件（LEECs）到等价评价条件（EECs）等不同级别的等价性。基于这个概念，我们提出了一种跨学科的统一评价工程学，我们称之为“基准学”（benchmarkology）。

在这个定义框架内，评价基准由三个基本组成部分构成。第一个组成部分是利益相关者的评价需求，其中包括各种因素，例如风险函数 (Risk function) 用于评估与评价基准相关的潜在风险。此外，还考虑了评价结果的差异阈值（discrepancy threshold），用于确定评价结果可接受的偏差水平。评价置信度 (Evaluation confidence level) 和评价置信度区间 (Evaluation confidence interval) 在预测完美评价模型（EM）的参数方面起着关键作用，最后，还考虑了 EM 的评价成本以及评价所需的资源。通过考虑这些要素，评价基准基准可以有效地满足利益相关者的评价需求。

评价基准的第二个组成部分是 EC 配置和机制。这包括对评价基准效果至关重要的几个要素。首先，它涉及定义利益相关者在解决问题或任务时面临的问题或任务集合。此外，还包括一组等价问题或任务实例，这有助于确保评价过程的相关性。评价基准还考虑了算法或者类算法机制及其实例化，这在解决所定义的问题或任务方面起着重要作用。支撑系统也被考虑在内，它们提供必要的资源和环境。

同时，评价基准提供了配置关键独立变量的手段，以及用以消除可能影响评价结果的混杂变量 (confounding)。此外，评价基准提供了解决利益相关者的多样化评价需求的机制。例如，它确保了不同等级的 EC 等价性，用以确定不同评价基准实例可视为等价的程度。

通过考虑这些 EC 配置和机制，评价基准可以提供全面和标准化的方法来解决评价问题。

第三个组成部分是指标（metrics）和参考（Reference），包括量的定义、价值函数、综合评估指标、参考评价对象和参考评价结果。

在本文的后续章节中，我们将把这三个组成部分称为评价基准的完整组成部分。图 5显示了评价基准的三个基本组成部分。

4. 评价、计量和测试的差异 (The differences between evaluation, measurement and testing)

我们进一步阐明了评价、计量（measurement）和测试（testing）之间显著的差异。计量学是关于测量及其应用的科学。计量学的本质在于量的定义和相应的测量方法。测试预期 (Test oracle) 是一种用于验证个体或系统在特定执行期间是否正确执行的方法。测试是执行个体或系统以确定其（1）是否符合测试预期定义的特定行为（第一类）和/或（2）是否在测试预期定义的环境中正确运行（第二类）。

首先，重要的是要承认在更广泛的评价框架内，计量或测试是其中基础的组成部分。除了计量和测试外，评价还包含一系列步骤。这些步骤包括定义并施加评价条件于不同的评价对象，从而形成了评价模型或系统。一旦评价模型或系统建立，不同评价对象的影响可以通过计量和/或测试进行推断。此外，考虑每个计量的量存在固有真值 (True quantity)，我

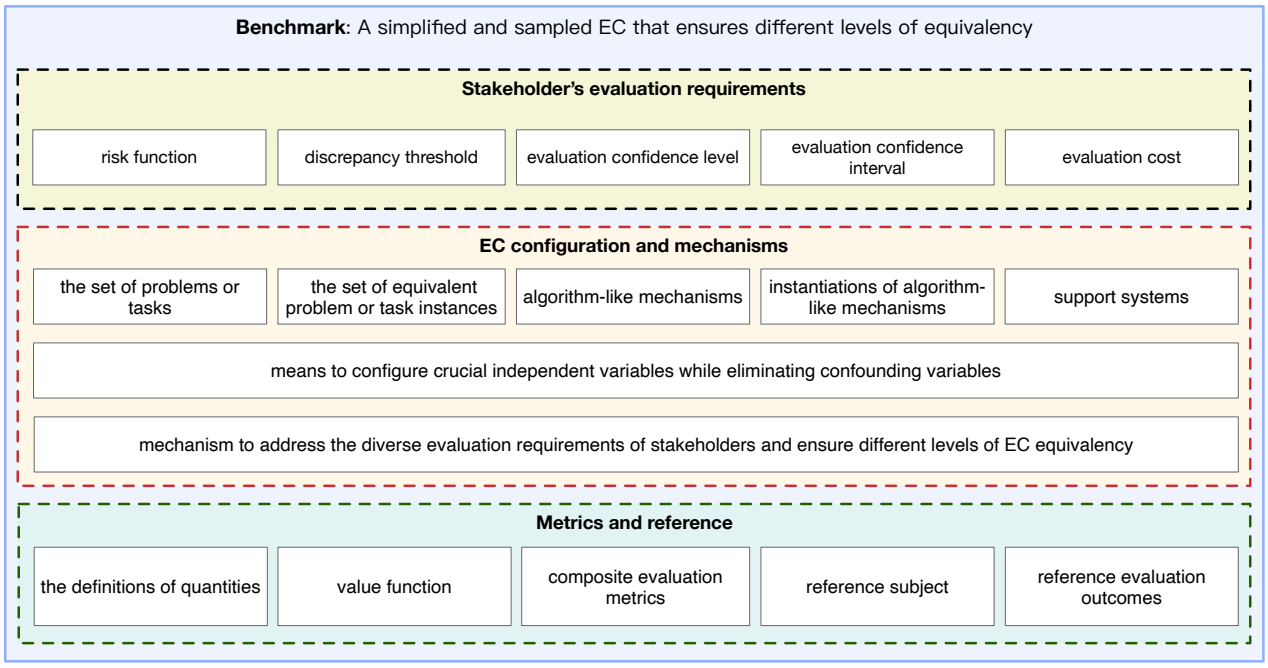

图 5: 评价基准 (Benchmark) 的三个基本组成部分。

们必须认识到计量结果是客观的。同样地，测试结果也具有客观性质，因为它们通常为每次测试提供正面或负面的结果。相反，评价结果具有一定程度的主观性，例如基于计量和/或测试数据定义的价值函数，我们已在第一个评价公理中讨论过它的主观性。基于上述原因，可以得出结论，计量学或测试学是评价学的基础知识体系。

5. 对现有评价和评价基准 (Benchmark) 实践的反思 (The reflections on state-of-the-art and state-of-the-practise benchmarks and evaluation)

为了进一步说明现有评价和评价基准实践的局限性，我们制作了图 8。通过检视该图，我们可以在评价学框架下了解现有最先进的评价研究和实践的不足之处。

显然，不同研究领域中评价的概念和术语缺乏共识。这种共识的缺乏往往导致混淆和误解，特别是当相同术语在不同学科中具有不同含义时。例如，“评价基准” (Benchmark) 一词在计算机科学、金融和商业学科中常被使用，但并没有正式定义。此外，即使在这些领域内，“评价基准”的定义也是模糊的并且容易引起误解。类似地，心理学可能使用“量表” (Scale) 一词作为类似于基准的概念，而社会科学和医学可能根本没有类似的概念。

认识到这一挑战，我们的工作旨在提出能够弥合这些学科差距的通用概念和术语。通过建立明确和标准化的定义，我们力求促进在不同学科之间共享对概念的理解，并推动不同研究领域之间的有效沟通与合作。

另一方面，目前鲜有研究讨论评价的本质，更不用说达成共识了。评价通常被错误地等同于计量 (measurement) 或测试 (testing)，而且没有明确的区分。例如，在计算机科学和心理学中，评价和计量经常被互换使用。同时，在测试（Testing）的背景下，目标是确定个体或系统是否与测试预期（Test oracle）定义的行为一致，评价通常与测试混淆。例如，根据 SPEC 术语，评价基准（benchmark）是指“用于比较一个计算机系统与其他系统性能的测试 (test) 或一组测试” [10, 12]。SPEC 是一个非常有影响力的评价基准组织。我们的工作揭示了评价的本质。

现有工作所提出的评价理论和方法往往是领域特定的，缺乏普遍适用的基础原则以及能跨越不同学科的评价方法。不同学科并不深入探讨评价的基本原则。相反，他们采用务实的方法，并优先考虑在

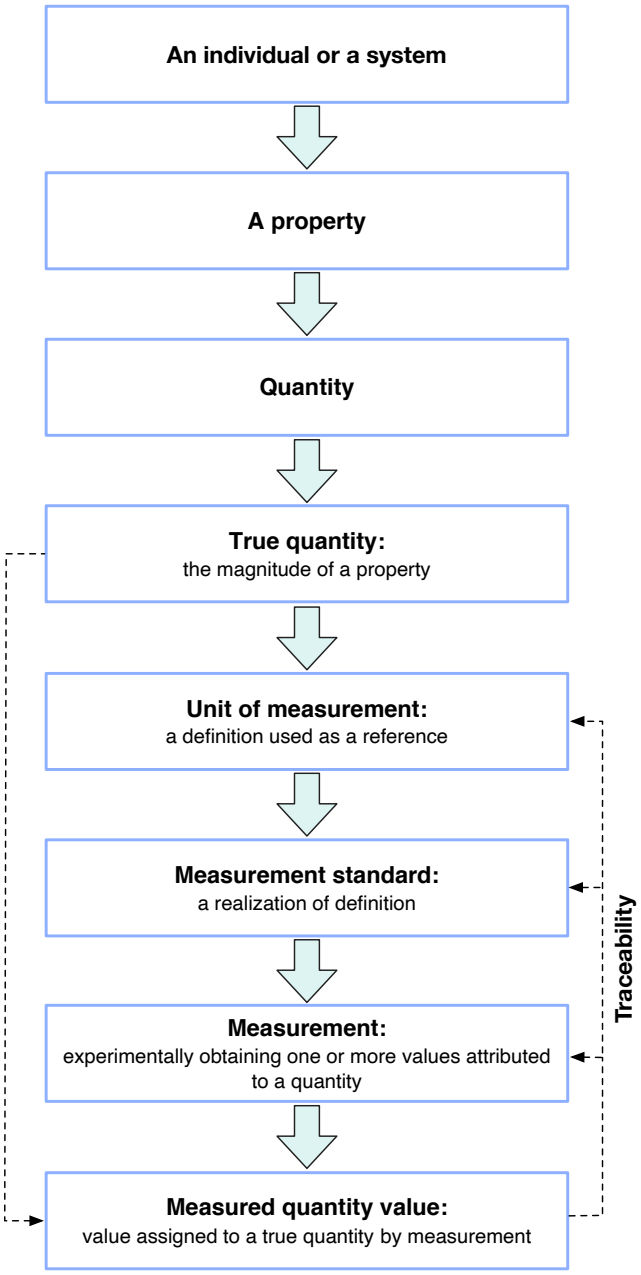

图 6: 简化的计量学 (Metrology) 概念框架。 [2, 8].

特定上下文中进行评价的指导方针。例如，在医学领域，重点主要在于消除特定研究组或队列内的混杂变量。最严格的理论基础可以在临床试验领域找到。例如，随机对照试验（RCT）技术用于排除混杂变量的影响。然而，这些方法或者技术缺乏通用的问题定义或者基础性的解决方案，不能充分考虑不同场景中评价模型（EMs）关键组件之间的复杂相互关系。

RCT 方法及其变体有两个严重的缺点。首先，缺乏严格的评价条件（EC）和等价评价条件（EEC）的层次化定义。EC 的变化可能引入混杂变量，从而影

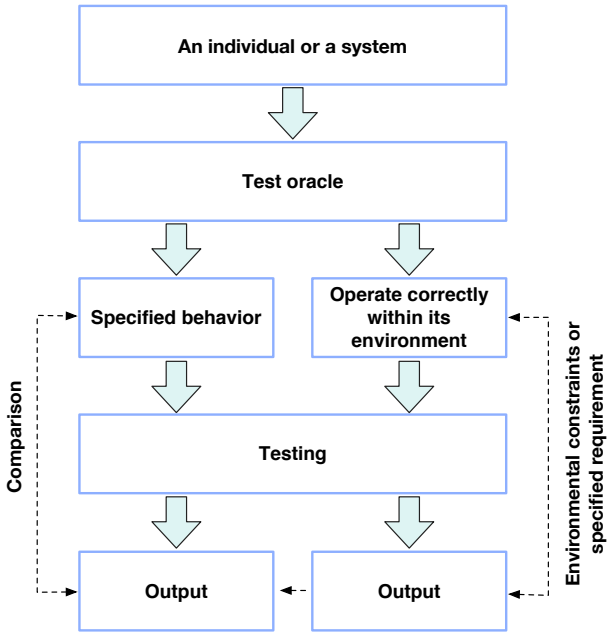

图 7: 简化的测试 (Testing) 概念框架。 [1, 14].

响评价结果并难以进行有意义的比较。在没有确保等价评价条件的情况下，期望评价结果之间能相互比较是一种幻想。其次，当涉及到研究复杂系统，如人类或实验动物，我们称之为支持系统时，RCT 方法及其变种可能难以建立一个参考评价模型（REM）。这种支持系统的特点是存在大量数目的独立变量，使得在受控实验设置中难以隔离和控制所有相关因素。因此，完全消除混杂变量并确保无偏差的评价结果变得非常具有挑战性。

在商业和金融领域，广泛使用了不同的观察研究方法（observational study methodologies），然而，观察研究甚至不是一种实验，它不能消除混杂变量并揭示因果关系（cause-and-effect relationships）。在商业学科中，标杆工程（Benchmarking）相当于评价学中定义的类算法机制的最先进实现和参考评价结果 (reference evaluation outcome)。在金融和教育学科中，评价基准 (Benchmark) 或指数（index）起着在观察研究中测量感兴趣的变量但不试图影响结果的参考评价结果的作用 [13]。

Rossi 等人提出了一个有价值的框架 [11]，用于评价社会科学领域的方法。然而，他们并没有提供一个可以应用于不同学科的通用理论。他们的局限在于仅狭隘地关注评价社会项目 (social program)，而没有发展出能够在复杂条件下评价其他对象的普遍理

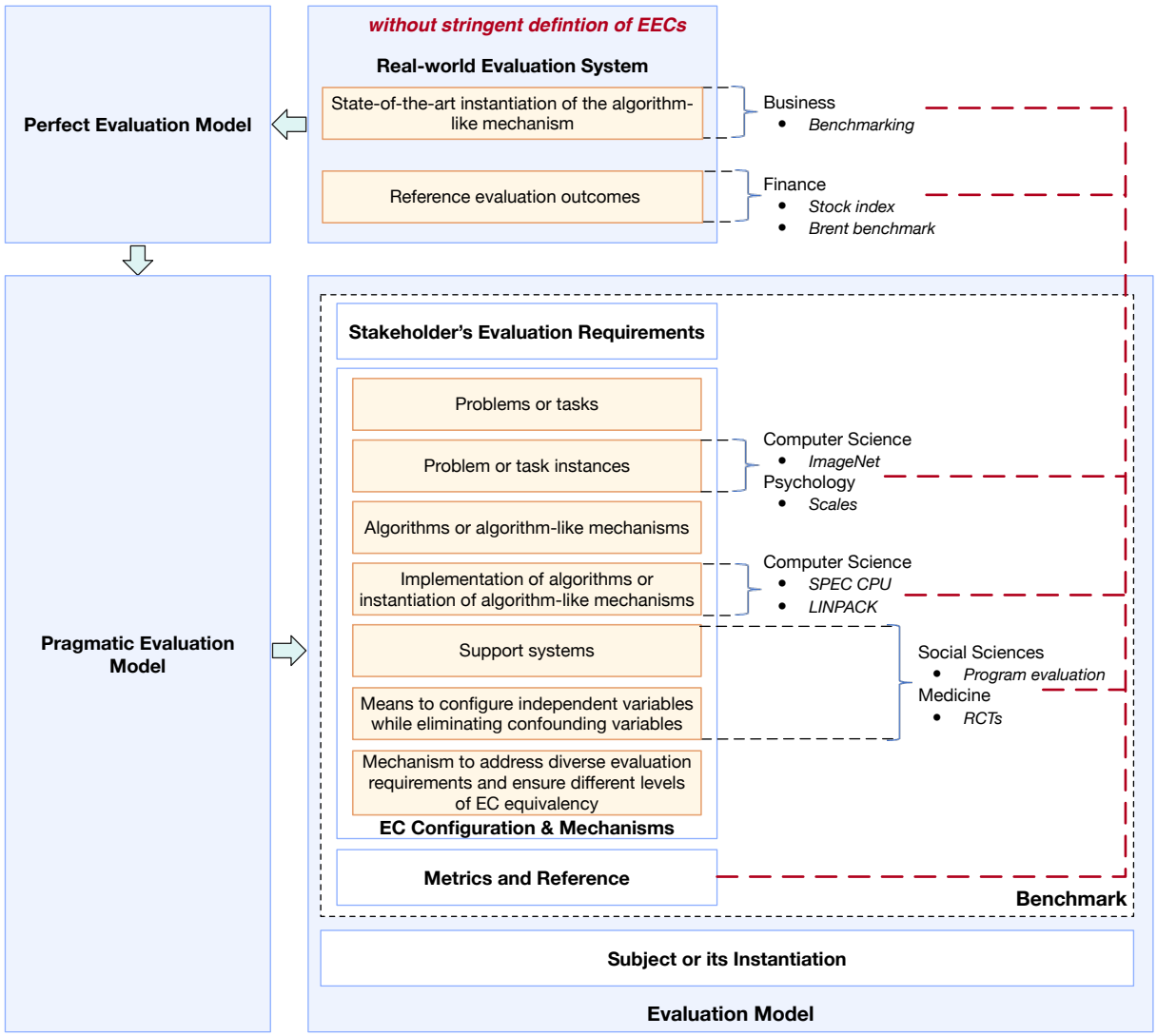

图 8: 在评价学理论框架下，对现有评价和评价基准 (Benchmark) 实践的反思。

论。Rossi 等人确实使用或开发了一些方法来隔离社会项目的影响，例如比较组设计（comparison group designs）和随机对照试验（RCT），但他们未能发展评价科学和工程的统一的基本原则和方法（underlying principles and methodology）。

在计算机科学领域，针对评价存在不同的观点和看法。例如，Hennessy 等人强调了评价基准的重要性 [6]，并将其定义为专门选择用于测量计算机性能的程序（programs specifically selected for measuring computer performance）。另一方面，John 等人编写了一本关于性能评价 (performance evaluation) 和基准评价 (Benchmarking) 的书 [7]，但没有提供这些概念的正式定义。Kounev 等人提出了评价基准的正式定义 [10]，即 “结合特定特征（如性能、可靠性或安全性），用于评价和比较系统或组件的工具及其方法”。ACM SIGMETRICS 小组 [3, 9] 认为性能评价（performance evaluation）是产生能够揭示计算机系统组件的执行频率和时间的数据，前提是有一套有序且明确定义的分析和定义步骤。

在心理学中，社会心理学家和人格心理学家经常使用量表 (scale)，如心理测量工具、测试或问卷调查（psychological inventories, tests, or questionnaires）来评价心理测量变量 [5]。虽然这些工具被广泛使用，但需要认识到它们依赖于虚拟评价和自我报告，这可能引入潜在的扭曲。为了克服这种限制，我们建议将评价条件实际地应用于评价对象，并辅以各种测量仪器来观察。这种方法旨在通过结合有形和可观察的数据，更客观和准确地评价各种方面，包括

态度、特质、自我概念、自我评价、信念、能力、动机、目标和社会认知（attitudes, traits, self-concept, self-evaluation, beliefs, abilities, motivations, goals, and social perceptions） [5]。

很多学科都提出了工程性的评价方法。然而，它们未能提供普遍 (universal) 的评价基准 (benchmark) 概念、理论、原则和方法。

例如，评价基准在金融、计算机科学和商业中被广泛使用，但其意义和实践却不一致。遗憾的是，已有的研究很少讨论能普遍适用于不同学科的评价基准的原则和方法（universal benchmark principles and methodologies）。Kounev 等人 [10] 为评价基准提供了全面的基础，包括指标、统计技术、实验设计等，但他们是从计算机科学的视角出发。

大多数最先进和最常用的评价基准研究和实践忽略了一个重要方面：利益相关者的评价需求。这种忽视导致未能考虑不同和多样的评价需求。例如，在评价中不将评价结果的差异控制在阈值范围内（enforce the discrepancy threshold in evaluation outcomes），也没有考虑评价置信度和置信区间等关键因素。因此，大多数处理器评价基准无法满足安全关键、任务关键和业务关键应用场景中的评价需求。

另一个问题是缺乏等价评价条件 (EEC) 或者最小等价评价条件 (LEEC) 等概念的的严格定义。大多数 CPU 或 AI（深度学习）评价基准，如 ImageNet，未能提供 EEC 或 LEEC 的明确定义。相反，它们直接跳转到标有真实值（Ground Truth）的特定数据集或者算法的实现，而没有给出充分的理由。此外，在某些情况下起着关键作用的支持系统被省略，而且没有解释简化评价基准的前提条件。此外，大多数方法未能讨论消除混杂因素的机制。这种忽视可能会在评价结果中引入偏差和不准确性。

毫不奇怪的是，在大多数评价基准的设计和实现过程中，对于复杂的评估机制 (mechanism) 和策略 (policy) 往往没有明确的讨论，未能解决一些重要的问题。例如，很少调研和表征真实世界评价系统 (ES)，很少涉及完美评价模型的设计和实现，很少关注从真实世界评价系统到评价模型的建模策略和过程，以及从完美评价条件到实用评价条件的采样策略和过程等重要方面。这一遗漏使得评价基准难以适应复杂的评价场景。

为了确保评价基准在复杂评价场景中的适用性和有效性，必须包含这些机制和策略。如果没有对真实世界评价系统的明确讨论，就很难建立一个能够捕捉真实世界评价系统的特征和利益相关者评价需求的评价条件。此外，在使用评价基准来估计真实世界系统的参数时，探索不同的采样 (sampling) 和建模 (modelling) 策略对于提升评价信心 (Evaluation confidence) 至关重要。通过仔细设计这些策略，我们可以在获得接近真值的评价结果和有效管理评价成本之间取得平衡。

目前有许多广泛使用的 AI（深度学习）评价基准。以 ImageNet 数据集为例 [4]，我们揭示了其局限性。首先，像 ImageNet 这样的特定 AI 评价基准不能追溯到明确的问题或任务的定义，而是以包含真实值的数据集的形式表现出来，这可能具有某些偏差。在其他评价场景中，我们也面临精确数学建模的挑战：例如对人体内化学和生物活动进行数学建模，或者对目标人群的社会动力学 (social dynamics) 进行数学建模。其次，评价基准依赖于一个未经验证的假设，即真实世界中的数据分布在很大程度上与所收集的数据集相一致。第三，在实际应用中，我们使用样本——特定评价基准——的统计数据 (statistics) 来推断总体 (population) 的参数 (parameter)。然而，我们不知道它们的置信度 (confidence level) 和置信区间 (confidence interval)。

References

[1] Baresi, L., Young, M., 2001. Test oracles .

[2] BiPM, I., IFCC, I., IUPAC, I., ISO, O., 2012. The international vocabulary of metrology—basic and general concepts and associated terms (vim). JCGM 200, 2012.

[3] Browne, J.C., 1975. An analysis of measurement procedures for computer systems. ACM SIGMETRICS Performance Evaluation Review 4, 29–32.

[4] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., . Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, IEEE. pp. 248–255.

[5] Furr, M., 2011. Scale construction and psychometrics for social and personality psychology. Scale Construction and Psychometrics for Social and Personality Psychology , 1–160.

[6] Hennessy, J.L., Patterson, D.A., 2011. Computer architecture: a quantitative approach. Elsevier.

[7] John, L.K., Eeckhout, L., 2018. Performance evaluation and benchmarking. CRC Press.

[8] Kacker, R.N., 2021. On quantity, value, unit, and other

terms in the jcgm international vocabulary of metrology. Measurement Science and Technology 32, 125015.

[9] Knudson, M.E., 1985. A performance measurement and system evaluation project plan proposal. ACM SIGMETRICS Performance Evaluation Review 13, 20–31.

[10] Kounev, S., Lange, K.D., Von Kistowski, J., 2020. Systems Benchmarking. Springer.

[11] Rossi, P.H., Lipsey, M.W., Henry, G.T., 2018. Evaluation: A systematic approach. Sage publications.

[12] SPEC, 2023. SPEC Glossary. https://www.spec.org/spec/glossary.

[13] Starnes, D.S., Yates, D., Moore, D.S., 2010. The practice of statistics. Macmillan.

[14] Whittaker, J.A., 2000. What is software testing? And why is it so hard? IEEE software 17, 70–79.

[15] Zhan, J., Wang, L., Gao, W., Li, H., Wang, C., Huang, Y., Li, Y., Yang, Z., Kang, G., Luo, C., et al., 2024. Evaluatology: The science and engineering of evaluation. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4, 100162.

詹剑锋博士是中国科学院计算技术研究所（ICT，CAS）研究员和中国科学院大学（UCAS）岗位教授，同时担任中国科学院计算技术研究所分布式系统研究中心主任。他于 1996 年和 1999 年分别获得西南交通大学的土木工程学士学位和固体力学硕士学位，并于 2002 年获得中国科学院软件研究所和中国科学院大学的计算机科学博士学位。他的研究领域涵盖从芯片到系统以及基准测试的各个方面，其共同主线是基准测试、设计、实现和优化。他在将学术研究转化为先进技术并应用于通用生产系统方面做了一些有效的工作。他的团队拥有的 35 项专利以及多项技术创新和研究成果已被转移至工业界。在过去的二十年中，他指导了近 90 多名研究生、博士后和工程师。詹剑锋教授创建了国际测试委员会（BenchCouncil）并担任了主席，创办了国际期刊 BenchCouncil Transactions on Benchmarks, Standards and Evaluation，并与 Tony Hey 教授共同担任联合主编。从 2018 年至 2022 年，他担任了 IEEE TPDS 的副编辑。他于 2006 年获得国家科技进步二等奖，2005 年获得中国科学院杰出成就奖，2013 年获得 IISWC 最佳论文奖，2024 年获得 Frontier of Computer Science 期刊的 Test of time paper award。