

A Simple Version of Evaluatology: The Science and Engineering of Evaluation

Jianfeng Zhan^{a,b,c,*}, Lei Wang^{b,a,c} and Wanling Gao^{b,a,c}

^aThe International Open Benchmark Council

^bICT, Chinese Academy of Sciences, Beijing, China

^cUniversity of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:

Evaluation
Benchmark
Scale
Index
Evaluation condition
Evaluation model
Evaluation system
Evaluation standard
Equivalent Evaluation condition
Least Equivalent Evaluation condition
Evaluatology
Benchmarkology

ABSTRACT

Evaluation is a crucial aspect of human existence and plays a vital role in each field. However, it is often approached in an empirical and ad-hoc manner, lacking consensus on universal concepts, terminologies, theories, and methodologies. This lack of agreement has significant consequences. This article aims to formally introduce the discipline of evaluatology, which encompasses the science and engineering of evaluation. The science of evaluation addresses the fundamental question: "Does any evaluation outcome possess a true value?" The engineering of evaluation tackles the challenge of minimizing costs while satisfying the evaluation requirements of stakeholders. To address the above challenges, we propose a universal framework for evaluation, encompassing concepts, terminologies, theories, and methodologies that can be applied across various disciplines, if not all disciplines.

This is a concise summary of Evaluatology [15]. For a more comprehensive understanding, please refer to the original article available at <https://www.sciencedirect.com/science/article/pii/S2772485924000140>. If you wish to cite this work, kindly cite the original article.

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. Introduction

Evaluation is a crucial aspect of human existence and plays a vital role in each field. However, it is often approached in an empirical and ad-hoc manner, lacking consensus on universal concepts, terminologies, theories, and methodologies. This lack of agreement has significant consequences. Even within computer sciences and engineering, it is not uncommon for evaluators to generate greatly divergent evaluation outcomes for the same individual or system under scrutiny, which we refer to as the *subject*. These discrepancies can range from significant variations to the extent of yielding contradictory qualitative conclusions. An example of this phenomenon can be observed when using multiple widely recognized CPU benchmark suites to assess the performance of the same processor. This often leads to greatly divergent evaluation outcomes that are incomparable across different benchmarks. A fundamental question arises in the realm of evaluation: "Does any evaluation outcome possess a true value?" Such circumstances give rise to valid concerns surrounding the reliability, effectiveness, and efficiency of these approaches when appraising the subject that is critical to safety, missions, and businesses.

For the first time, this article aims to formally introduce the discipline of evaluatology, which encompasses the science and engineering of evaluation. The science of evaluation addresses the fundamental question: "Does any evalua-

tion outcome possess a true quantitative measure?" The engineering of evaluation tackles the challenge of minimizing costs while satisfying the evaluation requirements of stakeholders. To address the above challenges, we propose a universal framework for evaluation, encompassing concepts, terminologies, theories, and methodologies that can be applied across various disciplines, if not all disciplines. Fig. 1 presents the universal concepts, theories, and methodologies in Evaluatology.


2. The science of evaluation

2.1. The essence of evaluation

The challenge in evaluation arises from the inherent fact that evaluating a subject in isolation falls short of meeting the expectations of stakeholders. Instead, it is crucial to apply a well-defined evaluation condition (EC) that reflects the stakeholders' concerns or interests. By doing so, evaluation can be viewed as an experiment that intentionally applies a well-defined EC to a subject. This process allows for the creation of an evaluation system or model. By measuring and/or testing this evaluation system or model, we can infer the impact of different subjects.

2.2. Five evaluation axioms

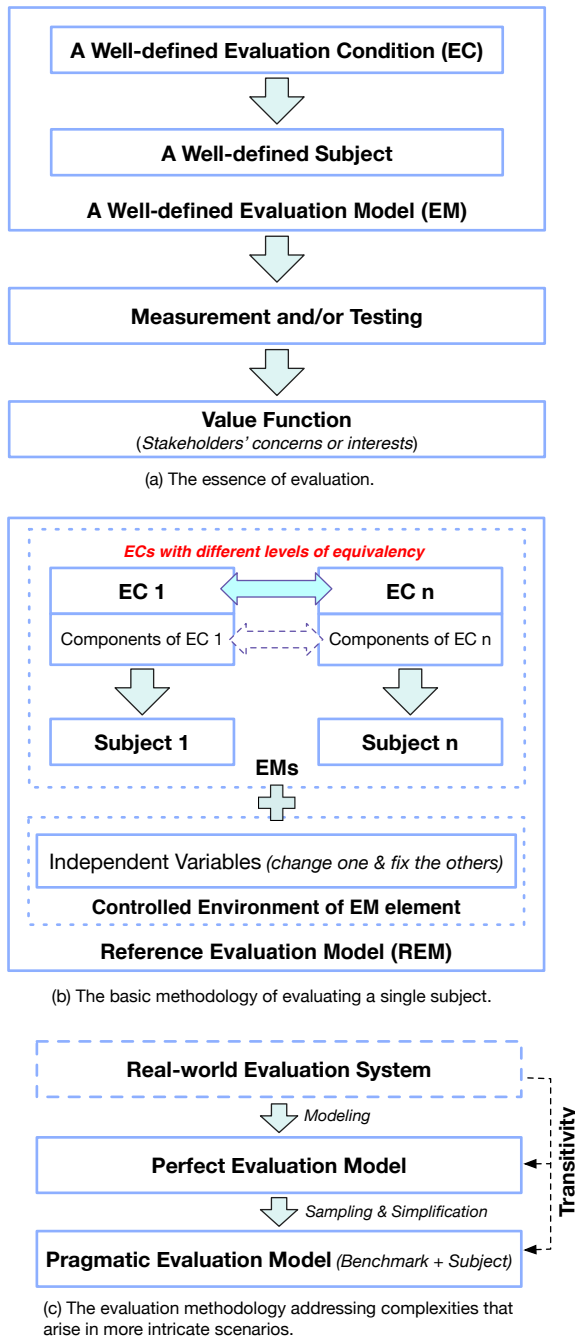
Derived from the essence of evaluation, we propose five axioms focusing on key aspects of evaluation outcomes as the foundational evaluation theory. These axioms serve as the bedrock upon which we build universal evaluation theories and methodologies.

 jianfengzhan.benchcouncil@gmail.com (J. Zhan)

 www.zhanjianfeng.org (J. Zhan)

ORCID(s):

Figure 1: The universal concepts, theories, and methodologies in evaluatology.



The Axiom of the Essence of Composite Evaluation Metrics declares that the essence of the composite evaluation metric either carries inherent physical significance or is solely dictated by the value function.

The Axiom of True Evaluation Outcomes declares that when a well-defined EC is applied to a well-defined subject, its evaluation outcomes, including its quantities and composite evaluation metrics, possess true values.

The Axiom of Evaluation Traceability declares that for the same subject, the divergence in the evaluation outcomes can be attributed to disparities in ECs, thereby establishing

evaluation traceability.

The Axiom of Comparable Evaluation Outcomes declares when each well-defined subject is equipped with equivalent ECs, their evaluation outcomes are comparable.

The Axiom of Consistent Evaluation Outcomes asserts that when a well-defined subject is evaluated using different samples from a population of ECs, their evaluation outcomes consistently converge towards the true evaluation outcomes of the population of ECs.

2.3. Basic evaluation theory

Based on the five evaluation axioms, we present the universal evaluation theories.

2.3.1. The hierarchical definition of an EC

A well-defined EC serves as a prerequisite for meaningful comparisons and analyses of the subjects. We propose a universal hierarchical definition of an EC and identify five primary components of an EC from the top to the bottom. We start defining an EC from the problems or task spaces that these stakeholders face and need to address with the following two reasons. 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 face and resolve, which provide a reliable means to define an EC. Second, utilizing the same problem or task can ensure the comparability of evaluation outcomes.

While the problem or task itself serves as the foundation for the evaluation process, it cannot solely serve as the evaluation itself because the problem or task is often abstract and requires further instantiation to determine its specific parameters. The second component is the set of a collective of equivalent problem or task instances, each of which is instantiated from the element of the first component. Different from the first component, an equivalent problem or task instance is specific and could serve as the evaluation directly. After a problem or task instance is proposed, it is necessary to figure out a solution. The third component consists of the algorithms or algorithm-like mechanisms, each of which provides the solution to a specific problem or task instance. An algorithm-like mechanism refers to a process that operates in a manner similar to an algorithm. The fourth component encompasses the implementation of an algorithm or instantiation of an algorithm-like mechanism. The fifth component is support systems that provide necessary resources and environments.

2.3.2. The establishment of EECs or LEECs

In the process of evaluating subjects, it is of utmost importance to prioritize the use of the equivalent ECs (EECs) across diverse subjects. This means that in order to establish two EECs, it is crucial to ensure that the corresponding components within the same layer of the two ECs are equivalent. By maintaining equivalency at each layer, we can ensure fair and unbiased evaluations, enabling meaningful comparisons and assessments between different subjects.

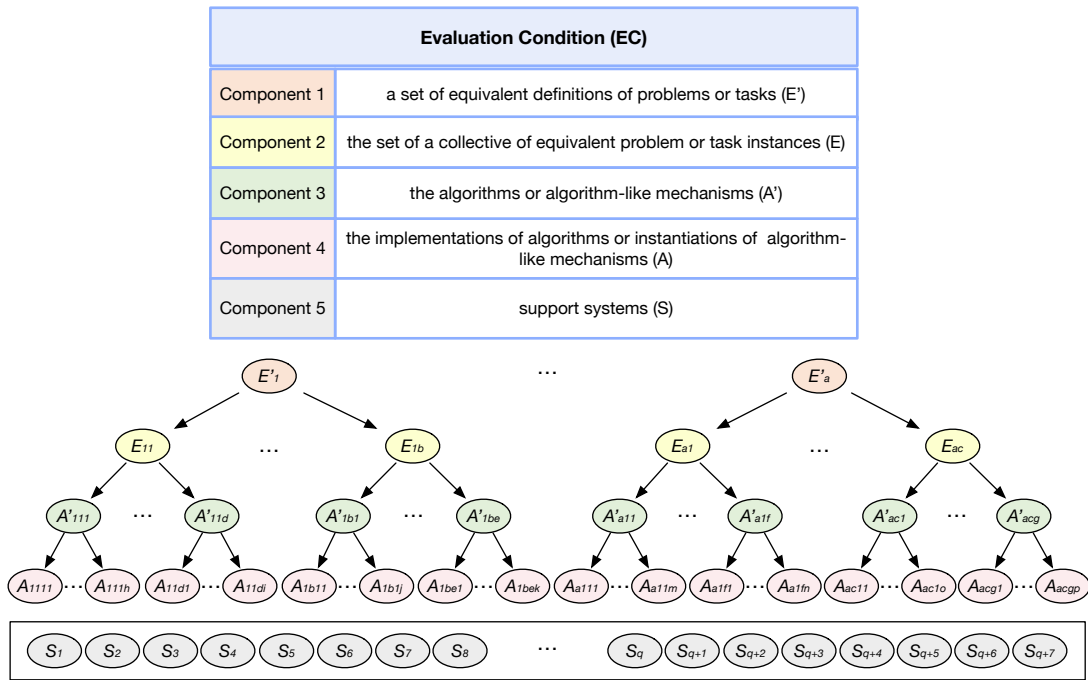


Figure 2: The Hierarchical Definition of an EC.

In certain cases, achieving complete equivalence between two ECs at all levels can be a challenging or even unattainable task. In such cases, we propose a minimum requirement of ensuring uniformity in the most essential components of the two ECs, which we refer to as the least equivalent evaluation conditions (LEECs). We propose the establishment of LEECs at the levels of the first and second top components of ECs.

To establish the LEECs, we identify the most governing component within an EC that must exhibit equivalency. This component, known as the evaluation standard, plays a crucial role in defining the LEECs. An evaluation standard should embody the characteristics that are solvable, definite, and equivalent (abbreviated as SDE). We propose the establishment of an evaluation standard at the level of the definition of an individual problem or task instance.

2.3.3. The establishment of an REM

We apply ECs with different levels of equivalency to diverse subjects to constitute EMs. An EM element refers to a specific point within the EM state space, and each EM element may have many independent variables. To eliminate confounding, we propose a new concept named a reference evaluation model (REM). An REM mandates that each element of an EM change only one independent variable at a time while keeping the other independent variables as controls. Subsequently, we utilize the measurement and/or testing to gauge the functioning of the REM. Finally, from the amassed measurement and testing data of the evaluation systems, we then deduce the cause-effect impacts of the different subjects.

2.4. Universal evaluation methodology in complex scenarios

Addressing the complexities that arise in more intricate scenarios, we reveal that the key to effective and efficient evaluations in various complex scenarios lies in the establishment of a series of EMs that maintain transitivity. In the full original version [15], we have formally defined what is transitivity in a mathematical form.

In real-world settings, we refer to the entire population of real-world systems that are used to evaluate specific subjects as the real-world evaluation system (ES). Assuming no safety concerns are present, the real-world ES serves as a prime candidate for creating an optimal evaluation environment, enabling the assessment of diverse subjects. However, there are several significant obstacles to consider, i.e., the presence of numerous confounding, the challenges of establishing an REM, prohibitive evaluation costs resulting from the huge state spaces, multiple irrelevant concurrent problems or tasks taking place, and the inclination to exhibit bias towards certain clusters within the EC state space.

We posit the existence of a perfect EM that replicates the real-world ES with utmost fidelity. A perfect EM eliminates irrelevant problems or tasks, has the capability to thoroughly explore and comprehend the entire spectrum of possibilities of an EC, and facilitates the establishment of REMs. However, the perfect EM possesses huge state space, entails a vast number of independent variables, and hence results in prohibitive evaluation costs. To address this challenge, it is crucial to propose a pragmatic EM that simplifies the perfect EM in two ways: reducing the number of independent variables that have negligible effect and sampling from the extensive state space. A pragmatic EM provides a means to

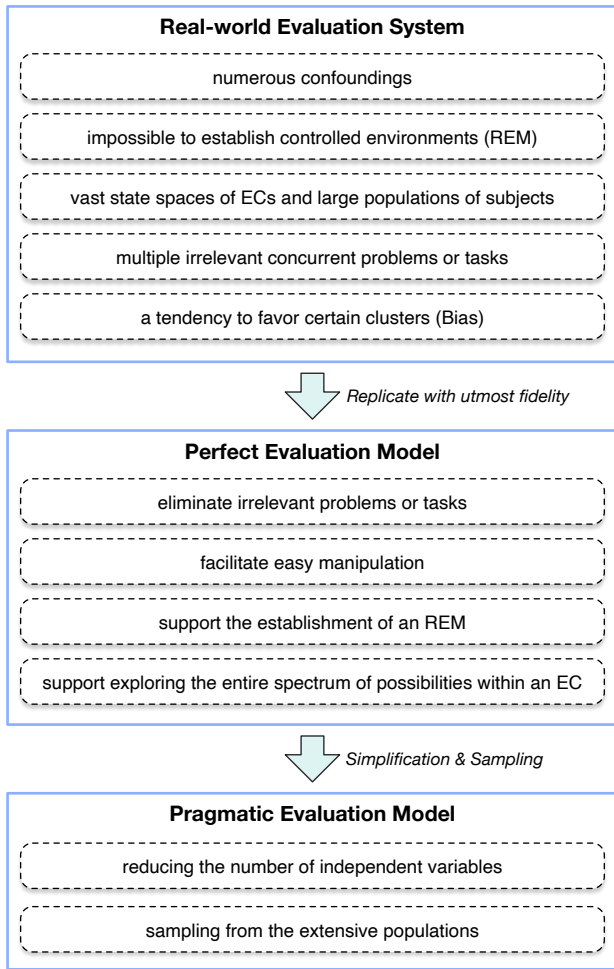


Figure 3: Universal evaluation methodology in complex scenarios.

estimate the parameters of the real-world ES.

2.5. Fundamental issues in evaluatology

We put forth four fundamental issues in the evaluations and formally formulate the problems mathematically in the full original version [15]:

First, how to ensure the transitivity of EMs is the most fundamental issue in building EMs in complex scenarios, from a real-world evaluation system to perfect EMs and pragmatic EMs.

Second, how to perform a cost-efficient evaluation with controlled discrepancies is the most important engineering issue in implementing evaluation processes. That is how to strike a balance between ensuring the discrepancy threshold of the evaluation outcomes and managing the associated costs.

Third, how to ensure evaluation traceability is a multi-faceted issue that requires the application of both scientific and engineering principles. It involves attributing any divergence in evaluation outcomes to disparities in the underlying ECs, thereby establishing clear and transparent traceability.

Fourth, how to connect and correlate evaluation standards across every discipline is the grand unified theory of

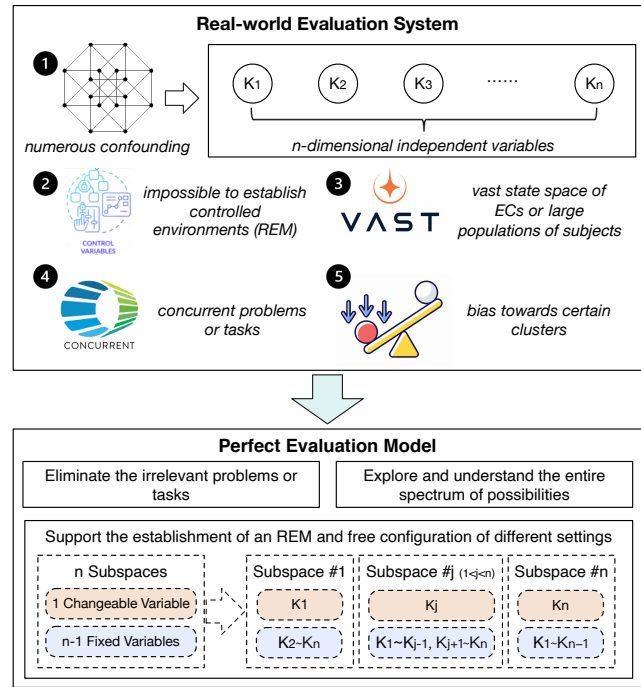


Figure 4: A perfect EM resembles a real-world ES.

evaluatology, allowing for a thorough exploration of evaluation-related matters. The evaluation standard serves as a fundamental pillar within any evaluation model. By establishing connections between evaluation standards across various disciplines, we have the potential to construct a comprehensive framework encompassing evaluation issues in all fields.

3. Benchmarkology: the engineering of evaluation

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 a *simplified and sampled EC, specifically a pragmatic EC, that ensures different levels of equivalency, ranging from LEECs to EECs*. Based on this concept, we propose a benchmark-based universal engineering of evaluation across different disciplines, which we aptly term “benchmarkology.”

Within the framework of this definition, a benchmark comprises three essential constituents. 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. The evaluation confidence level and evaluation confidence interval play a crucial role in predicting the parameter of a perfect EM. Lastly, the evaluation cost of EM is taken into account, and the resources required for conducting the evaluation are assessed. By considering these elements, the benchmark can effectively address the

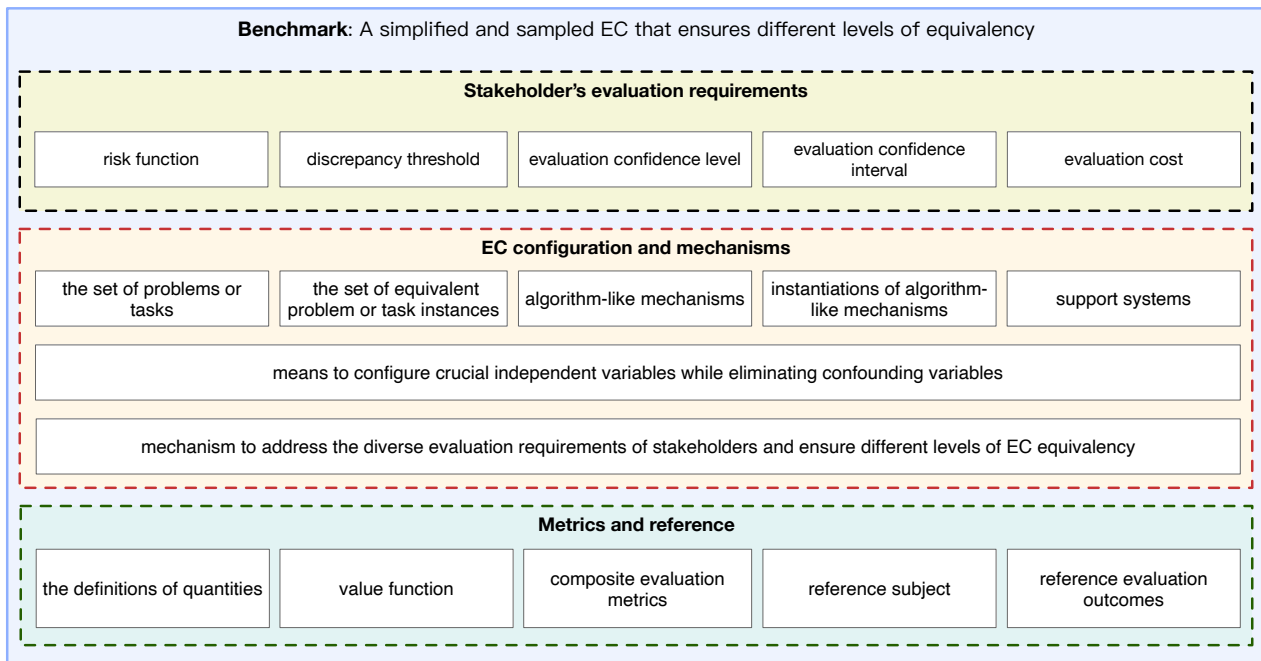


Figure 5: A benchmark comprises three essential constituents.

evaluation requirements of stakeholders.

The second constituent of the benchmark framework is the *EC configuration and mechanisms*. This includes several elements crucial for the benchmark's effectiveness. Firstly, it involves defining the set of problems or tasks that the stakeholders face when addressing them. Additionally, it encompasses the set of equivalent problem or task instances, which helps ensure specificity in the evaluation process. The benchmark also considers algorithm-like mechanisms and their instantiations, which play a significant role in solving the defined problems or tasks. The support systems, which provide necessary resources and environments, are also taken into account.

Moreover, the benchmark provides the means to configure crucial independent variables while eliminating confounding variables that could potentially impact the evaluation outcomes. Also, the benchmark provides the mechanism to address the diverse evaluation requirements of stakeholders. For example, it ensures different levels of EC equivalency, determining the extent to which different benchmark instances can be considered equivalent.

By considering these EC configurations and mechanisms, the benchmark can provide a comprehensive and standardized approach to different evaluation issues.

The third constituent is the *metrics and reference*, including the definitions of quantities, the value function, composite evaluation metrics, the reference subject, and the reference evaluation outcomes.

In the subsequent sections of this article, we will refer to these three constituents as the complete constituents of a benchmark. Figure 5 shows the three essential constituents of a benchmark.

4. The differences between evaluation, measurement and testing

We elucidate the marked disparity between evaluation, measurement, and testing.

Metrology is the science of measurement and its applications. The essence of metrology lies in quantities and their corresponding measurements.

A test oracle is a method used to verify whether an individual or system being tested has performed correctly during a specific execution. Testing is the process of executing an individual or system to determine whether it (1) conforms to the specified behavior defined by the test oracles (the first category) and/or (2) operates correctly within its intended environment as defined by the test oracles (the second category).

First and foremost, it is important to acknowledge that measurement or testing serves as a preliminary constituent within the broader framework of evaluation. In addition to measurement and testing, an evaluation encompasses a series of steps. These steps involve defining and applying evaluation conditions to a diverse range of subjects, which ultimately leads to the creation of an evaluation model or system. Once the evaluation model or system is established, the impacts of different subjects can be inferred through the process of measuring and/or testing.

Furthermore, it is crucial to recognize that the measurement results are of an objective nature, assuming the existence of an inherent true value for each measured quantity. Similarly, testing results also possess an objective nature as they typically yield either a positive or negative outcome for

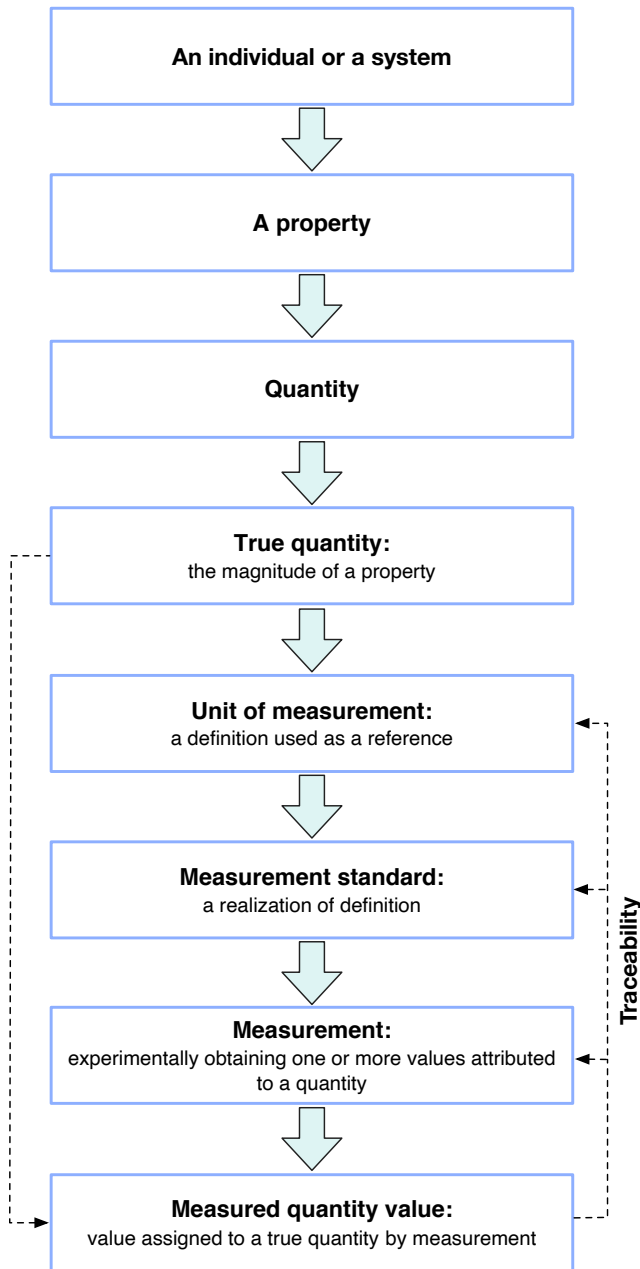


Figure 6: A simplified yet systematic conceptual framework for metrology [2, 8].

each test conducted.

Conversely, evaluation results possess a certain degree of subjectivity, such as the formulation of value functions based on the underlying measurement and/or testing data, which we have discussed in the first evaluation axiom. By virtue of the aforementioned reasons, we can assert that metrology or testing serves as but one foundational aspect in the realm of evaluations.

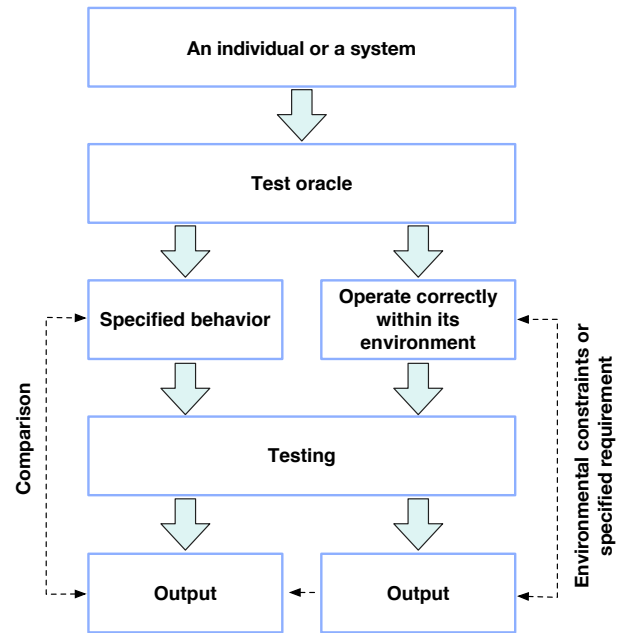


Figure 7: A simplified yet systematic conceptual framework for testing [1, 14].

5. The reflections on state-of-the-art and state-of-the-practise benchmarks and evaluation

To further illustrate the limitations of existing evaluation and benchmark practices, we present Fig. 8, which showcases these shortcomings within the evaluatology framework. By examining this figure, we can gain a clearer understanding of the areas where state-of-the-art and state-of-the-practice evaluation and benchmarks fall short.

It is evident that a lack of consensus exists regarding concepts and terminologies across different areas of study. This lack of consensus often leads to confusion and misinterpretation, especially when the same terms are used in different disciplines with varying meanings. For example, the term “benchmark” is commonly employed in computer science, finance, and business disciplines but without a formal definition. Moreover, even within these fields, the definition of “benchmark” can be vague and subject to interpretation. In contrast, psychology may use the term “scale” as a concept similar to benchmark, while social science and medicine may not have an analogous concept at all.

Recognizing this challenge, our work has aimed to propose universal concepts and terminologies that can bridge these disciplinary gaps. By establishing clear and standardized definitions, we seek to promote a shared understanding and facilitate effective communication and collaboration across different areas of study.

Few works discuss the essence of evaluation, let alone reaching a consensus on it. Evaluation is often mistakenly equated with measurement or testing without clear differentiation. For instance, in computer science and psychology, evaluation and measurement are often used interchangeably.

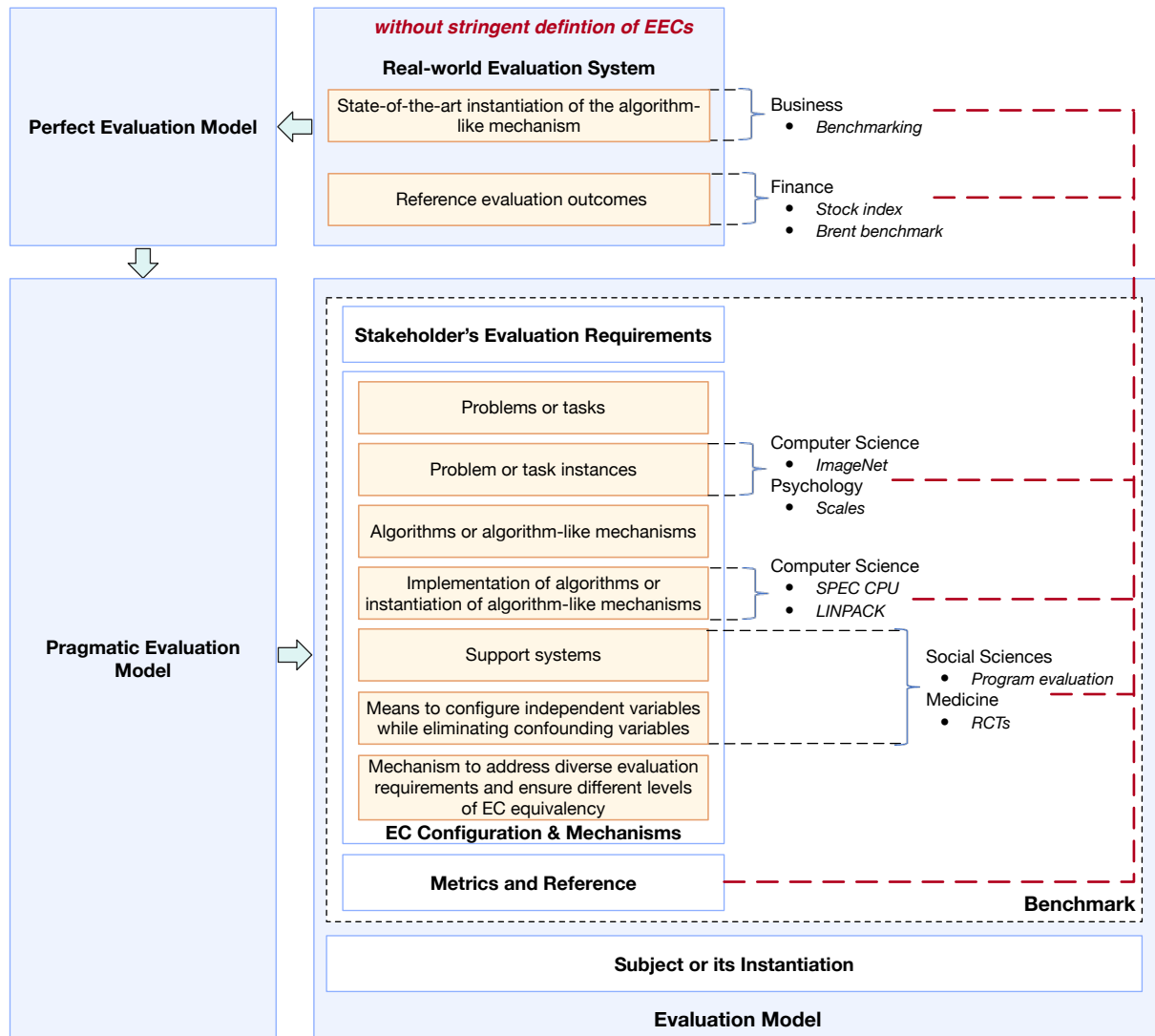


Figure 8: The reflections on state-of-the-art and state-of-the-practice benchmarks and evaluation are based on the science and engineering of evaluation.

In the context of testing, where the goal is to determine whether an individual or a system aligns with the expected behavior defined by test oracles, evaluation is often conflated with testing. For instance, according to the SPEC terminology, a benchmark refers to “a test, or set of tests, designed to compare the performance of one computer system against the performance of others” [10, 12]. SPEC is a highly influential benchmark organization. Our work has revealed the essence of the evaluation.

The proposed evaluation theories and methodologies are often domain-specific, with a lack of universally applicable foundational principles and evaluation methodologies that transcend diverse disciplines. Different disciplines do not delve into the underlying principles of evaluation. Instead, they adopt a pragmatic approach and prioritize guidelines for conducting evaluations within specific contexts. For instance, in the medical discipline, the focus is primarily on eliminating confounding variables within the specific groups or cohorts being studied. In the business discipline, efforts

are concentrated on searching the state of the practice.

The most rigorous theoretical foundation can be found in the field of clinical trials. For instance, Randomized Controlled Trial (RCT) techniques are employed to rule out the effect of confounding variables. However, there is a lack of universal problem formulations or fundamental solutions that fully consider the intricate interactions among the key components of EMs in diverse scenarios.

There are two serious drawbacks to the RCT methodology and its variants. Firstly, there is a lack of a stringent hierarchical definition of EC and EECs. The variations in ECs can introduce confounding that may affect the results and make meaningful comparisons difficult. Without ensuring EECs, it becomes an illusion to expect comparable evaluation outcomes. Secondly, when it comes to studying complex systems such as human beings or experimental animals, which we refer to as support systems, the RCT methodology and its variants may struggle to establish an REM. This kind of support system is characterized by a mul-

titude of independent variables, making it difficult to isolate and control all relevant factors in a controlled experimental setting. Consequently, it becomes challenging to eliminate confounding variables and ensure unbiased evaluation outcomes completely.

In the realms of business and finance, different observational study methodologies are widely used. An observational study is not even an experiment. Certainly, it cannot eliminate confounding variables and reveal the cause-and-effect relationships. In the business discipline, benchmarking assumes the state-of-the-art instantiation of the algorithm-like mechanism and the reference evaluation outcomes. In finance and education disciplines, benchmarks or indexes assume the role of reference evaluation outcomes in an observational study that measures variables of interest but does not attempt to influence the response [13].

Rossi et al. [11] propose a valuable framework for evaluating methodologies in the field of social science. However, they do not provide a universal theory that can be applied to different disciplines. Their limitations stem from their narrow focus on assessing social programs without developing a generalized theory for evaluating other subjects in complex conditions.

Rossi et al. indeed utilized or developed some approaches to isolate the social programs' impacts, e.g., comparison group designs and randomized controlled trials (RCT), but they failed to explicitly state the underlying principles and methodology for universal science and engineering of evaluation.

Within the computer science field, there are varying viewpoints and perspectives. For example, Hennessy et al. [6] highlight the significance of benchmarks and define them as programs specifically selected for measuring computer performance. On the other hand, John et al. [7] compile a book on performance evaluation and benchmarking without providing formal definitions for these concepts. Kounev et al. [10] 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 [3, 9] 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.

In psychology, social and personality psychologists often utilize scales, such as psychological inventories, tests, or questionnaires, to assess psychometric variables [5]. 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, we suggest implementing a physical application of an EC to the subjects, 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 [5], by incorporating tangible and observable data.

Various disciplines have proposed engineering approaches to evaluations. However, they fail to provide universal benchmark concepts, theories, principles, and methodologies.

For instance, benchmarks are commonly utilized in finance, computer science, and business, albeit with inconsistent meanings and practices. Regrettably, there have been limited discussions in previous works regarding universal benchmark principles and methodologies that can be applied across different disciplines. From a computer science standpoint, Kounev et al. [10] provide a comprehensive foundation for benchmarking, including metrics, statistical techniques, experimental design, and more.

Most state-of-the-art and state-of-the-practice benchmarks overlook an essential aspect: the stakeholders' evaluation requirements. This oversight leads to a failure to consider different and diverse evaluation requirements. For instance, they do not enforce the discrepancy threshold in evaluation outcomes, nor do they consider evaluation confidence level and confidence interval, among other crucial factors. As a result, most CPU benchmarks are ill-equipped to meet the evaluation requirements in scenarios involving safety-critical, mission-critical, and business-critical applications.

Another issue is the lack of a stringent definition for similar concepts, such as an EEC or LEEC. For example, most CPU or AI (deep learning) benchmarks, like ImageNet, fail to provide a clear definition of an EEC or LEEC. Instead, they jump directly into a specific dataset labeled with the ground truth or the implementation of algorithms without proper justification. Additionally, the support system, which plays a crucial role in some cases, is omitted without any explanation of the condition of simplifying the benchmarks. Furthermore, most of the methodologies fail to discuss the confounding elimination mechanism. This oversight can potentially introduce bias and inaccuracies in the evaluation outcomes.

Not surprisingly, the intricate evaluation mechanisms and policies are not explicitly discussed in the design and implementation of most benchmarks. For instance, it fails to address important aspects such as investigating and characterizing real-world ES, the design and implementation of a perfect EM, the modeling policy and procedure from a real-world ES to an EM, and the sampling policy and procedure from a perfect EC to a pragmatic EC. This omission makes it difficult for the benchmark to adapt to intricate evaluation scenarios.

It is crucial to include these mechanisms and policies to ensure the benchmark's applicability and effectiveness in complex evaluation scenarios. Without explicit discussion of the real-world ES, it is difficult to establish an EC that captures the characteristics and requirements of real-world evaluations. Furthermore, exploring different sampling and modeling policies is essential to gain the confidence of the evaluation community in using the benchmark for inferring parameters of real-world ES. By carefully designing these policies, we can strike a balance between achieving high accuracy in evaluation outcomes and managing the associated evaluation costs.

There are many widely used AI (deep learning) benchmarks. Taking the ImageNet dataset as an illustrative example [4], we reveal their limitations. Firstly, a specific AI benchmark like ImageNet cannot be traced back to an explicit formulation of a problem or task and instead manifests itself in the form of a dataset containing ground truth, which may possess certain biases. In other scenarios, we also encounter challenges in identifying a precise mathematical function that accurately models the chemical and biological activities within the human body or the social dynamics within the target population. Secondly, the benchmark relies on an unverified assumption that the data distribution within the real world closely aligns with that of the collected dataset to a considerable extent. Thirdly, in real-world applications, we use the statistic of a sample – a specific benchmark – to infer the parameters of the entire population. However, we do not know their confidence levels and confidence intervals.

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., Ye, H., Dai, S., Zhang, Z., 2024. Evaluatology: The science and engineering of evaluation. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4, 100162.

Jiaotong University in 1996 and 1999 and his Ph.D. in Computer Science from the Institute of Software, CAS, and UCAS in 2002. His research areas span from Chips and systems to Benchmarks. A common thread is benchmarking, designing, implementing, and optimizing diverse systems. He has made substantial and effective efforts to transfer his academic research into advanced technology to impact general-purpose production systems. Several technical innovations and research results, including 35 patents from his team, have been adopted in benchmarks, operating systems, and cluster and cloud system software with direct contributions to advancing parallel and distributed systems in China or worldwide. Over the past two decades, he has supervised over ninety graduate students, post-doctors, and engineers. Dr. Jianfeng Zhan is the founder and chairman of BenchCouncil. He also holds the role of Co-EIC of BenchCouncil Transactions on Benchmark, Standards and Evaluations, alongside Prof. Tony Hey. Dr. Zhan has served as an Associate Editor for IEEE TPDS (IEEE Transactions on Parallel and Distributed Systems) from 2018 to 2022. In recognition of his exceptional contributions, he has been honored with several prestigious awards. These include the second-class Chinese National Technology Promotion Prize in 2006, the Distinguished Achievement Award of the Chinese Academy of Sciences in 2005, the IISWC Best Paper Award in 2013, and the Test of Time Paper Award from the Journal of Frontier of Computer Science.



Dr. Jianfeng Zhan is a Full Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and University of Chinese Academy of Sciences (UCAS), the director of the Research Center for Distributed Systems, ICT, CAS. He received his B.E. in Civil Engineering and MSc in Solid Mechanics from Southwest