BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4 (2024) 100175

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Editorial

A short summary of evaluatology: The science and engineering of evaluation

Jianfeng Zhan *

The International Open Benchmark Council, DE, USA
ICT, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Evaluation
Benchmark
Evaluation system
Evaluation condition
Equivalent evaluation condition
Evaluatology

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 short summary of Evaluatology (Zhan et al., 2024). The objective of this revised version is to alleviate the readers' burden caused by the length of the original text. Compared to the original version (Zhan et al., 2024), this revised edition clarifies various concepts like evaluation systems and conditions and streamlines the concept system by eliminating the evaluation model concept. It rectifies errors, rephrases fundamental evaluation issues, and incorporates a case study on CPU evaluation (Wang et al., 2024). For a more comprehensive understanding, please refer to the original article (Zhan et al., 2024). 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. The motivation: why it is essential to establish the science and engineering of evaluation

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[1] 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 the industry-standard CPU benchmark suite SPEC CPU2017 to assess the performance of the same processor [4]. Wang et al. [4] used SPEC CPU2017 to evaluate the same X86 processor, an Intel Xeon Gold 5120T, adhering to SPEC's procedures and rules. In the rest of this article, we use the same CPU evaluation experiment. Nonetheless, across various SPEC CPU2017 configurations with alterations in compiler flags and the number of copies/threads, the best and worst outcomes exhibit a notable difference of 86 times. Under the SPEC CPU2017 recommended configuration,[2] just 12 out of 43 SPEC CPU2017 workloads achieved the best performance among the varying configurations. From a measurement or metrology perspective, each procedure and

* Correspondence to: The International Open Benchmark Council, DE, USA.
E-mail address: jianfengzhan.benchcouncil@gmail.com.
URL: https://www.zhanjianfeng.org.

[1] This footnote is quoted from [1]. An individual can be defined as an object described by a given data set. A system is an interacting or interdependent group of individuals, whether of the same or different kinds, forming a unified whole [2,3].

[2] The SPEC CPU2017 recommended configuration sets the compiler flag to '-O3' and the number of threads/copies to the maximum number of hardware threads supported by the CPU.

https://doi.org/10.1016/j.tbench.2024.100175

Available online 12 September 2024

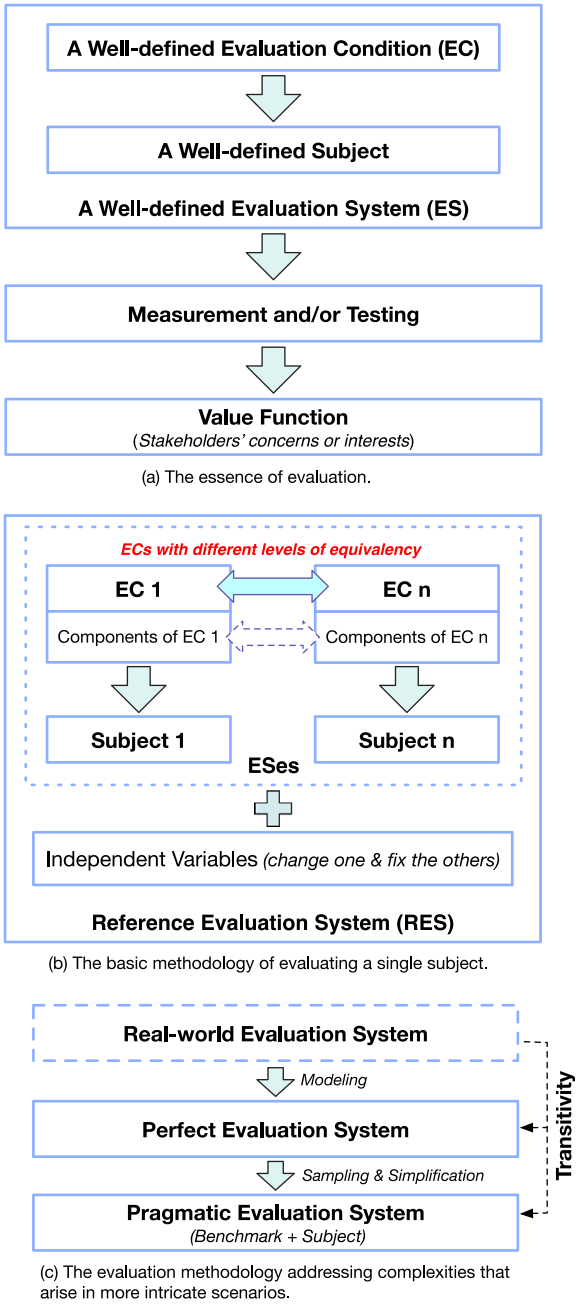

Fig. 1. The universal concepts, theories, and methodologies in evaluatology.

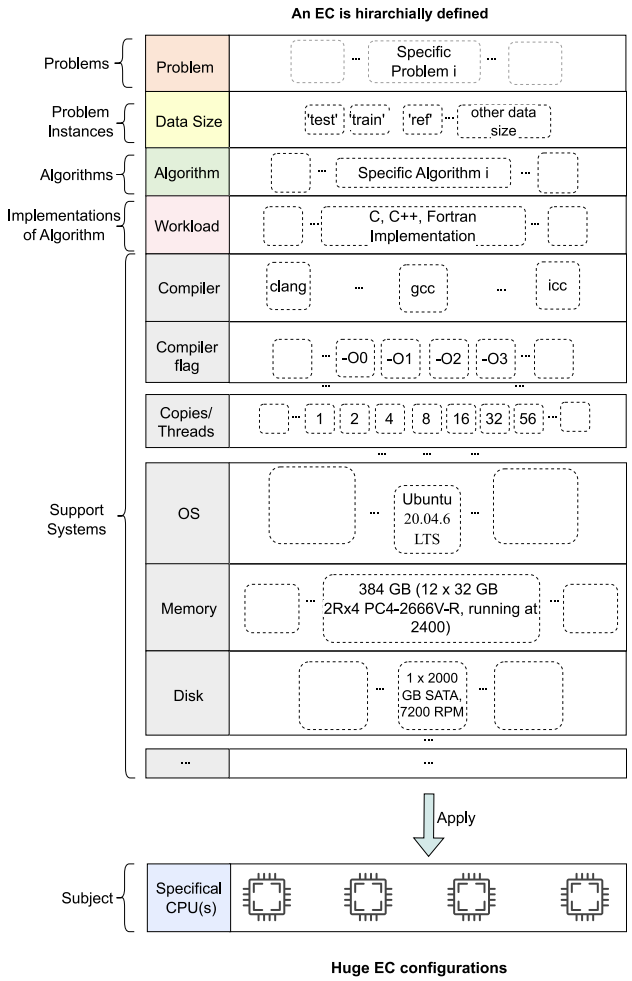

Fig. 2. In the context of CPU evaluation, a well-defined EC possesses huge configurations built on the basis of SPEC CPU2017, an industry-standard CPU benchmark suite. With the permissions of the authors of [4].

the obtained quantity is valid and correct. Nevertheless, the result under a particular configuration can be misleading if the distribution of outcomes across different configurations is not taken into account.

In this context, two fundamental questions naturally emerge: "What is the distinction between evaluation and measurement? Does any evaluation outcome possess a true value?" Such circumstances give rise to valid concerns surrounding these approaches' reliability, effectiveness, and efficiency 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 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. 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 create a minimal and well-defined evaluation system (ES) that satisfies the evaluation requirements of stakeholders. Providing the context to evaluate the subject, an ES is a minimum system consisting of the subject and other individuals or systems that are crucial in guaranteeing independent operation and addressing the concerns or interests of the subject's stakeholders.

In other words, evaluation can be seen as an experiment that deliberately applies a well-defined evaluation condition (EC) to a subject to create an ES. Building on the previous discussion, literally, an EC can be understood as the ES with the subject removed. We formally define EC as the minimal context that is crucial in guaranteeing independent operation and addressing the concerns or interests of the subject's stakeholders for evaluating the subject.

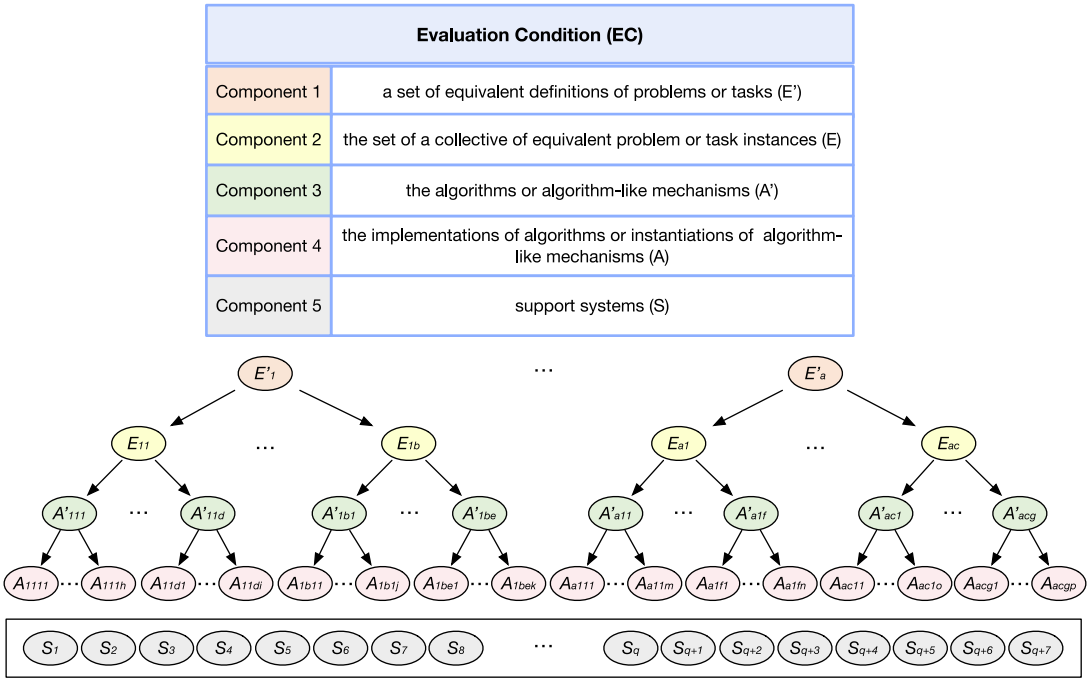

Fig. 3. The hierarchical definition of an EC.

A well-defined EC possesses huge EC configurations. For example, Fig. 2 presents an EC example built on the basis of SPEC CPU2017, an industry-standard CPU benchmark suite.

In one word, *evaluation is the process of inferring the impact of subjects indirectly within ES that cater to the requirements of stakeholders, relying on objective measurements and/or testing of the latter.*

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.

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

The Axiom of True Evaluation Outcomes states that when a well-defined EC is applied to a well-defined subject, the evaluation outcomes possess true values, representing a distribution of value across different EC configurations.

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 under different configuration samples of a well-defined EC, their evaluation outcomes consistently converge towards the true evaluation outcomes.

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. As shown in Fig. 3, 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 tasks that these stakeholders face and need to address with the following two reasons. First, the relevant stakeholders' concerns and interests are at the evaluation's core. 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. Secondly, employing the same problem or task provides the necessary but not sufficient method to ensure the comparability of evaluation outcomes. While the problem or task serves as the foundation for the evaluation process, it cannot solely serve as the evaluation itself because it is often abstract and requires further instantiation to determine its specific parameters.

The second component is the set of problem or task instances, each of which is instantiated from a problem or task. Different from the first component, a problem or task instance is specific and could serve as the evaluation directly. After a problem or task 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 problem or task. An algorithm-like mechanism refers to a process or abstract 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, which tackles problem or task instances. The fifth component is support systems that provide necessary resources and environments.

2.3.2. The establishment of EECs

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 guarantee that evaluation results are not influenced by confounding variables in ECs, allowing for meaningful comparisons and assessments across different subjects.

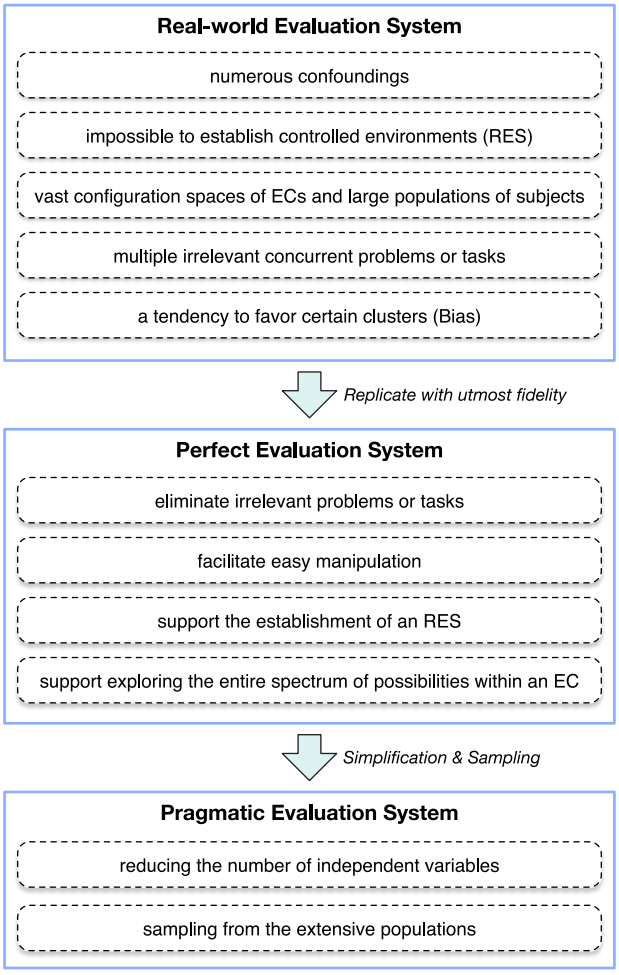

Fig. 4. Universal evaluation methodology in complex scenarios.

We take the same CPU evaluation example to demonstrate why it is essential to guarantee EECs. In the SPEC CPU2017 workloads, the outcomes vary significantly, ranging from 1.95 to 242.06, across different configurations of SPEC CPU2017. For instance, the workload named *541.leela_r* shows a 242.06-fold difference in outcomes for the Intel Xeon Gold 5120T between a configuration with the '-O3' compiler flag and 56 copies versus another configuration with the '-O0' compiler flag and 1 copy. These significant disparities in evaluation outcomes for the same processor highlight the importance and necessity of EECs.

2.3.3. The establishment of a reference ES

We apply ECs to diverse subjects to constitute ESes. An ES configuration refers to a specific point within the ES configuration space, and each ES configuration has many independent variables. There is a subtle difference between an EC configuration and an ES configuration, as the subject itself also may possess many independent variables.

We propose a new concept named a reference ES (RES) to address confounding variables. An RES mandates that each ES configuration changes only one independent variable at a time, maintaining the other variables as controls. Subsequently, we utilize the measurement and/or testing to gauge the functioning of the RES. Finally, from the amassed measurement and testing data of the ESes, we deduce the cause–effect impacts of the independent variable that we modify.

Similarly, we can define the concept of reference EC (REC). An REC mandates that each EC configuration changes only one independent variable at a time, maintaining the other variables as controls.

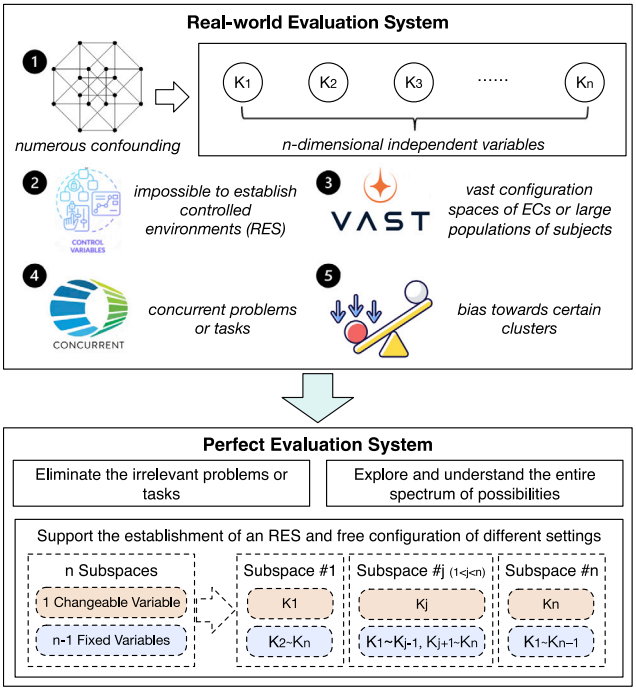

Fig. 5. A perfect ES resembles a real-world ES.

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 ESes that maintain transitivity (see Fig. 4). In the full original version [1], we have formally defined what transitivity is in a mathematical form.

In real-world settings, we refer to the minimal real-world systems that are used to evaluate specific subjects as the real-world 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 variables, the challenges of establishing an RES, prohibitive evaluation costs resulting from the huge configuration spaces, multiple irrelevant concurrent problems or tasks taking place, and the inclination to exhibit bias towards certain clusters within the ES configuration space.

We posit the existence of a perfect ES that replicates the real-world ES with utmost fidelity (see Fig. 5). A perfect ES eliminates irrelevant problems or tasks, has the capability to thoroughly explore and comprehend the entire spectrum of possibilities of an ES, and facilitates the establishment of an RES. However, the perfect ES possesses huge configuration 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 ES that simplifies the perfect ES in two ways: reducing the number of independent variables that have negligible effect and sampling from the extensive configuration space. A pragmatic ES provides a means to estimate the parameters of the real-world ES.

Literally, a real-world, perfect, or pragmatic EC can be understood as the corresponding ES without the subject included.

2.5. Four fundamental issues in evaluatology

We put forth four fundamental issues in the discipline of evaluatology.

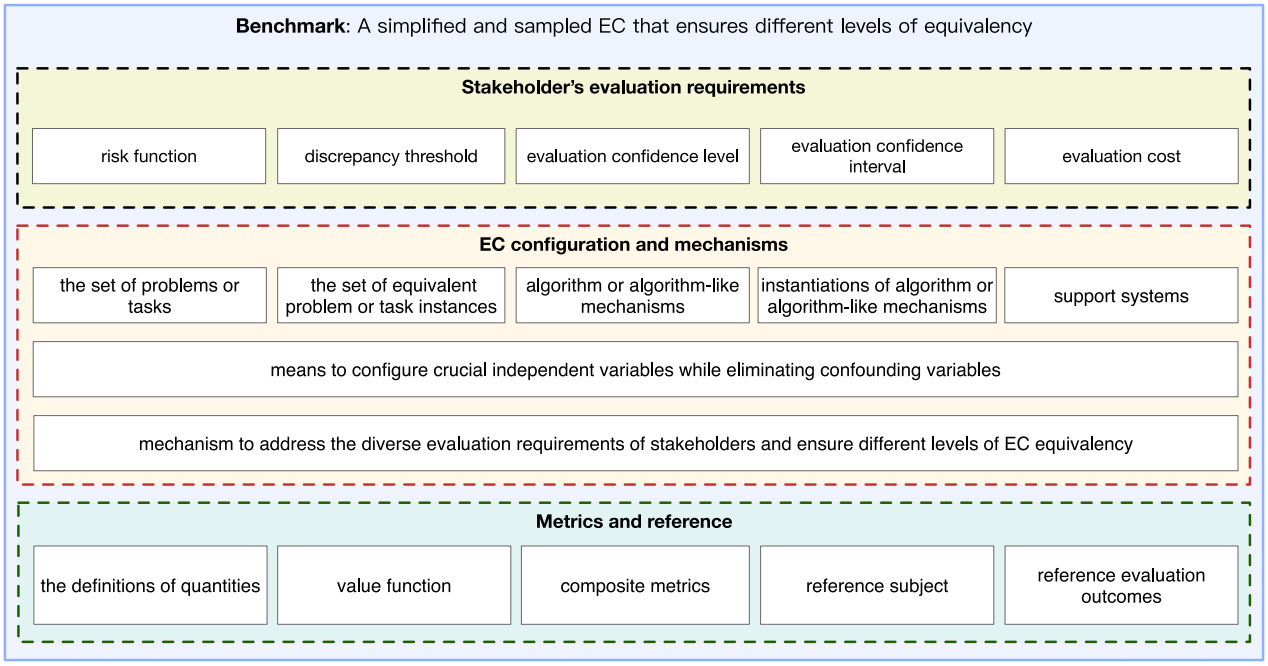

Fig. 6. A benchmark comprises three essential constituents.

First and foremost, establishing the EC that can yield true and consistent evaluation outcomes for the same subject stands as the cornerstone of evaluatology. Moreover, ensuring the transitivity of ECs is equally crucial in complex scenarios, transitioning from a real-world EC to perfect and pragmatic ECs.

Secondly, ensuring the evaluation outcomes are comparable for different subjects stands as another cornerstone of evaluatology. Given the formidable task of ensuring EECs and the potential multitude of independent variables within an EC and a subject, mitigating the adverse impacts of confounding variables poses a significant challenge.

Thirdly, the critical engineering challenge in implementing evaluation processes is determining how to conduct a cost-effective evaluation while maintaining controlled outcome discrepancies. That is how to strike a balance between ensuring the discrepancy threshold of the evaluation outcomes and managing the associated costs.

Fourthly, how to ensure evaluation traceability is a multifaceted issue that requires the application of both scientific and engineering principles. It involves attributing any discrepancy in evaluation outcomes to disparities in the underlying ECs and subjects, thereby establishing clear and transparent traceability.

3. 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 *an EC. In reality, it could be a simplified and sampled EC, specifically a pragmatic EC, that ensures different levels of equivalency*. Based on this concept, we propose a benchmark-based universal engineering of evaluation across different disciplines.

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 ES. 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 *EC configuration and mechanisms*. This includes several elements crucial for the benchmark's effectiveness. Firstly, it involves defining the set of problems or tasks the stakeholders face when addressing. Additionally, it encompasses the set of problem or task instances, which helps ensure specificity in the evaluation process. The benchmark also considers algorithms or algorithm-like mechanisms, which play a significant role in solving the defined problems or tasks, and includes their instantiations. 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 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. Fig. 6 shows the three essential constituents of a benchmark.

4. A case study of CPU evaluation

Three fundamental steps are involved in our examination of CPU evaluation as a case study in evaluatology. The first step is to delineate the EC and the subject. The second step entails applying the well-defined EC to the subject to establish the ES. Finally, the third step focuses on attaining consistent and comparable evaluation outcomes.

In CPU evaluation, a specific CPU, a well-defined subject, includes components such as decoders, issue queues, arithmetic logic units (ALUs), branch predictors, reorder buffers (ROBs), and caches.

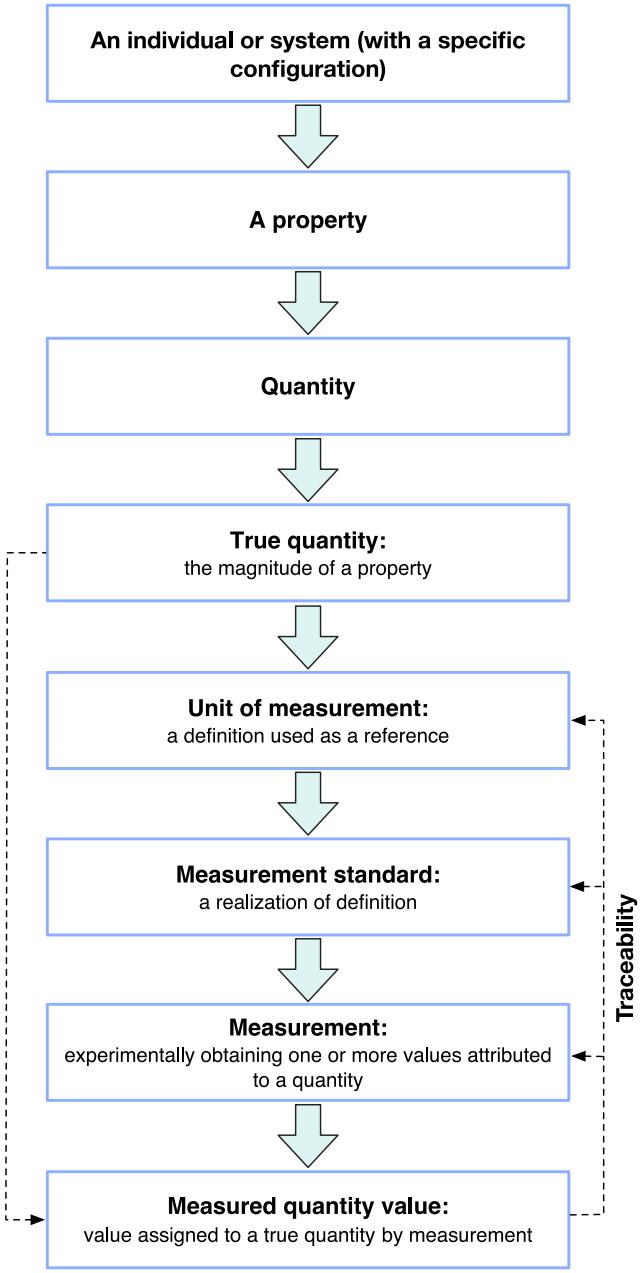

Fig. 7. A simplified yet systematic conceptual framework for metrology [5,6].

As shown in Fig. 2, a well-defined EC consists of five components. At the highest level are the problems that are of concern to stakeholders, such as arithmetic operations. The next level is the problem instances, which provide data variables and constraints of the problem, such as the arithmetic operations on two 32-bit unsigned integers n_1 and n_2. The third level is the algorithms, an abstract representation of the solution for the given problem, which can be described in natural language or pseudocode. The next level is algorithm implementations, which involve coding the algorithms in a specific programming language supported by the computer, such as a C or Java program. The final level is the support systems, which encompass all environments or configurations required to run a program on a specific CPU, including but not limited to the compilers, compiler flags, copies/threads, OSes, OS settings, memories, memory settings, disks, and disk settings.

When a well-defined EC is applied to a specific CPU, the true evaluation outcome emerges as a distribution of outcomes across various EC configurations. Under a specific EC configuration, the true outcome can

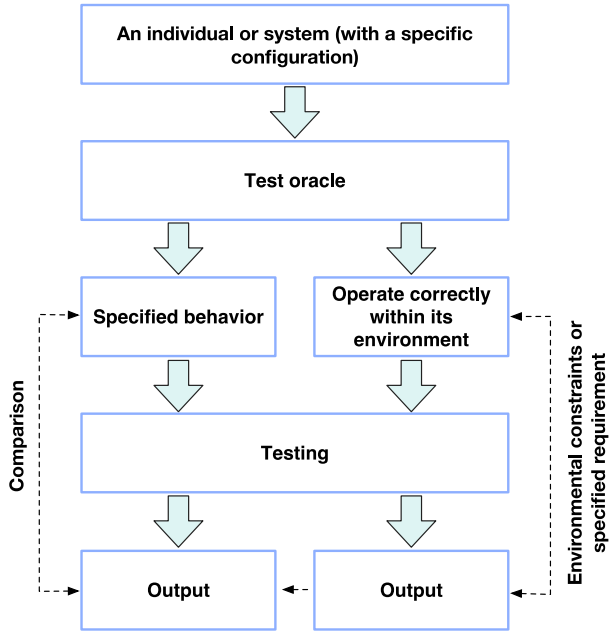

Fig. 8. A simplified yet systematic conceptual framework for testing [7,8].

be estimated by taking the mean or median of multiple experiments. Furthermore, under a sample of configurations of a well-defined EC, the true evaluation outcomes can be estimated using a confidence interval at a specified confidence level, with the sample mean serving as the estimate.

The evaluation outcomes of different CPUs are comparable under EECs. The EECs consist of identical problems, problem instances, algorithms, algorithm implementation, support systems containing the same compiler, compiler flag, copies/threads setting, OS, OS setting, memory, memory setting, disk, and disk setting in CPU evaluation. Any divergence in ECs will result in incomparable evaluation outcomes for different CPUs.

5. 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 (see Fig. 7).

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) (see Fig. 8).

First and foremost, measurement, testing, and evaluation focus on different issues within the same scenario, e.g., the CPU evaluation experiment. The primary focus of evaluation lies in defining an EC that can yield true and consistent evaluation outcomes for the same subject. Eliminating confounding variables within the EC to ensure that evaluation outcomes remain comparable across different subjects is another crucial issue. In this context, measurement and testing is a micro-level activity that addresses the previously mentioned issue within a particular EC configuration.

Secondly, measurement, testing, and evaluation serve distinct purposes within the same scenario. Evaluation focuses on the subject, while measurement or testing targets ES. Measurement or testing is carried

out directly on the ES, whereas evaluation is derived indirectly within the ES.

Thirdly, measurement, testing, and evaluation have different outcomes within the same scenario. The evaluation outcomes appear as the distribution of outcomes with respect to different EC configurations. However, testing and measurement outcomes appear as the distribution of outcomes with respect to the same EC configurations.

By virtue of the aforementioned reasons, we can assert that metrology or testing serves as but one foundational aspect in the realm of evaluations.

Acknowledgments

The author expresses gratitude to all the authors of "Evaluatology: The science and engineering of evaluation. BenchCouncil Transactions on Benchmarks, Standards, and Evaluations, 4(1), 100162". and acknowledges the contribution of Dr. Wanling Gao for refining all figures, and my Ph.D. student, Chenxi Wang, for providing the CPU evaluation example.

References

[1] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks Stand. Eval. 4 (1) (2024) 100162.
[2] System, 2024, https://www.merriam-webster.com/dictionary/system. (Accessed 6 February 2024).
[3] A. Backlund, The definition of system, Kybernetes 29 (4) (2000) 444–451.
[4] C. Wang, L. Wang, W. Gao, Y. Yang, Y. Zhou, J. Zhan, Achieving consistent and comparable CPU evaluation outcomes, 2024, Technical Report, International Open Benchmark Council.
[5] I. BiPM, I. IFCC, I. IUPAC, O. ISO, The international vocabulary of metrology—basic and general concepts and associated terms (VIM), JCGM 200 (2012) 2012.
[6] R.N. Kacker, On quantity, value, unit, and other terms in the JCGM international vocabulary of metrology, Meas. Sci. Technol. 32 (12) (2021) 125015.
[7] L. Baresi, M. Young, Test oracles, 2001.
[8] J.A. Whittaker, What is software testing? And why is it so hard? IEEE Softw. 17 (1) (2000) 70–79.

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 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 focus on evaluatology, evaluatology-based design automation, and optimization automation. His exceptional expertise is exemplified by his introduction to the discipline of evaluatology, an endeavor that encompasses the science and engineering of evaluation; within this discipline, his proposition of a universal framework for evaluation encompasses essential concepts, terminologies, theories, and methodologies for application across various disciplines. 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.