Benchmarking Without Constructs: A Measurement Theory Critique of MCP Evaluation Frameworks

Himaneesh Sompalle
Stonehill International School
himaneesh@emergence.ai

Abstract

The Model Context Protocol (MCP) has rapidly become a de facto standard for connecting large language models to external tools, prompting a wave of benchmarks—MCP-Bench, MCP-Universe, MCPMark, and others—aimed at evaluating agent competence in tool-use scenarios. We argue that this benchmark proliferation has outpaced construct definition: each framework implicitly encodes a different theory of what "MCP competence" means, yet none explicitly operationalizes the construct it purports to measure. Drawing on classical measurement theory (Cronbach and Meehl, 1955; Messick, 1995), we analyze three prominent MCP benchmarks along the axes of construct validity, evaluation reliability, and reproducibility. We find that the field faces a measurement crisis analogous to early psychometrics: instruments are being built before the constructs they target have been agreed upon. We propose that evaluation researchers adopt explicit construct operationalization, multi-trait validation protocols, and a structured interdisciplinary consensus process before further benchmark proliferation.

1 Introduction

The Model Context Protocol (Anthropic, 2024) has emerged as a transformative standard for connecting large language models (LLMs) to external data sources and tools. Within less than a year, this standardization has triggered a benchmark arms race: at least six distinct evaluation frameworks targeting MCP-mediated tool use appeared between April and September 2025, each claiming to measure "real-world" agent capabilities.

We focus on three prominent benchmarks—MCP-Bench (Wang et al., 2025), MCP-Universe (Luo et al., 2025), and MCPMark (Wu et al., 2025)—that collectively represent the methodological spectrum of current MCP evaluation. While each makes valuable contributions, we contend that the field is repeating a well-documented pattern from measurement science: building instruments before defining the constructs those instruments should measure (Cronbach and Meehl, 1955; Loevinger, 1957).

This matters beyond academic tidiness. As MCP-mediated agents enter production systems, evaluation frameworks will increasingly function as gatekeepers. If the constructs underlying these benchmarks remain implicit and divergent, the evaluation ecosystem risks producing scores that are internally consistent but externally meaningless—a problem that classical measurement theory terms the *jingle fallacy*: different tests measuring different things under the same label (Kelley, 1927; Block, 1995). Analogous crises have occurred in psychometrics (the "intelligence" debate), educational testing (the "reading comprehension" fragmentation), and clinical assessment (Borsboom, 2006)—and in each case, resolution required explicit construct negotiation before further instrument production.

Our contribution is twofold. First, we provide a structured measurement-theoretic analysis of three MCP benchmarks, showing that they target divergent implicit constructs. Second, we offer a concrete roadmap for interdisciplinary construct consensus, including the specific conditions under which a cross-benchmark empirical validation study becomes feasible.

2 Three Benchmarks, Three Implicit Constructs

We analyze MCP-Bench, MCP-Universe, and MCPMark along dimensions drawn from classical test theory (Lord and Novick, 1968; Embretson and Reise, 2000): the construct each benchmark implicitly targets, how it operationalizes that construct, and what evaluation methodology it employs. Table 1 summarizes the comparison.

Dimension	MCP-Bench	MCP-Universe	MCPMark
Implicit construct	Planning under ambiguity	Adaptive tool mastery	Operational depth
Operationalization	Multi-hop tasks with fuzzy instructions across 28 live servers, 250 tools	Long-horizon tasks across 6 domains, 11 servers with unfamiliar tool spaces	CRUD-diverse workflows in 5 real environments, 127 expert-curated tasks
Evaluation method	Rule-based checks + LLM-as-judge with prompt shuffling	Three-tier execution-based: format, static, and dynamic evaluators	Programmatic verification scripts with curated initial states
Reliability mechanism	Score averaging across shuffled prompts	Dynamic evaluators query live server state	Deterministic scripts; $pass^4$ metric over 4 runs
SOTA ceiling	Hierarchy across 20 LLMs; schema compliance converged at >95%	GPT-5 at 43.7%; significant long-context degradation	GPT-5-medium at 52.6% pass@1; 33.9% $pass^4$

Table 1: Comparative analysis of three MCP evaluation frameworks along measurement-theoretic dimensions.

MCP-Bench connects LLM agents to 28 live MCP servers spanning 250 tools. Its defining feature is *fuzzy instructions*: tasks do not name specific tools, forcing agents to infer which tools are appropriate from underspecified natural language. The implicit construct is *planning under ambiguity*—the ability to decompose a vague objective into a concrete tool-use trajectory. This construct has precedent in the planning literature (Bylander, 1994) and in task-oriented dialogue evaluation (Eric et al., 2020), where instruction underspecification is a recognized difficulty axis.

MCP-Universe spans 6 domains across 11 MCP servers with a deliberately broad tool inventory designed to test agents on options not encountered during training. The implicit construct is *adaptive tool mastery*—competence in the face of novelty and scale. This maps onto transfer learning evaluation paradigms (Hendrycks et al., 2021), where out-of-distribution generalization is the central concern.

MCPMark operates in 5 representative environments with 127 tasks emphasizing *CRUD-diverse workflows*—tasks requiring creating, reading, updating, and deleting resources rather than read-heavy retrieval. The implicit construct is *operational depth*—the ability to execute full write-path workflows with side effects. Evaluation uses programmatic verification with a $pass^4$ metric requiring success across four independent runs.

The three construct labels we assign are interpretive rather than empirically derived — a limitation we discuss in the Limitations section. We present them not as definitive taxonomic claims but as hypotheses to be tested through cross-benchmark validation.

3 A Measurement Theory Diagnosis

3.1 Construct Validity

None of the three benchmarks provides an explicit construct definition or nomological network (Cronbach and Meehl, 1955) for the capability it measures. This is not a flaw in any single benchmark—it reflects genuine dimensionality in the capability space—but without explicit construct boundaries, users cannot know what a high score *means* beyond performance on that specific task set.

The absence of convergent and discriminant validity testing (Campbell and Fiske, 1959) is particularly concerning. In established psychometric practice, a new instrument must demonstrate that it correlates with theoretically related measures (convergent validity) and diverges from theoretically unrelated ones (discriminant validity). Without such analysis, aggregate leaderboards that rank models across benchmarks commit a construct conflation error. The educational testing literature offers instructive precedent: decades of conflating "reading comprehension" subtests that measured different latent skills delayed both theoretical progress and practical diagnosis (Kintsch, 1994).

3.2 Evaluation Reliability

The three benchmarks occupy different points on the reliability–validity tradeoff (Cronbach, 1951):

- MCPMark's programmatic verification maximizes *internal consistency*: the same task, given the same initial state, yields a deterministic pass/fail. However, this constrains tasks to those where correctness is programmati-

cally verifiable, excluding partial credit and open-ended quality dimensions.

- MCP-Bench's LLM-as-judge approach enables richer evaluation of open-ended tasks but introduces *judge-model bias*: the evaluation inherits the judge LLM's own preferences. Prompt shuffling and score averaging mitigate but do not eliminate this variance (Zheng et al., 2023).

- MCP-Universe's dynamic evaluators query live server state, enabling evaluation of time-sensitive tasks but introducing *environmental non-determinism* that undermines test-retest reliability (Liao et al., 2021).

3.3 Reproducibility

Both MCP-Bench and MCP-Universe rely on live external servers. Server downtime, API changes, rate limiting, and data drift all threaten reproducibility. MCPMark partially addresses this through containerized environments with pinned server versions, but even this approach faces long-term maintenance challenges. None of the three benchmarks reports the monetary cost of a full evaluation run (Solaiman et al., 2025), despite cost being a central concern in evaluation infrastructure.

4 A Roadmap for Construct Consensus

Based on our analysis, we offer three concrete proposals—framed as provocations but accompanied by actionable steps.

Provocation 1: Declare your construct. Every MCP benchmark should begin with an explicit construct definition distinguishing it from adjacent capabilities. "Tool-use competence" is too broad; "the ability to decompose underspecified objectives into multi-hop tool-call trajectories without explicit tool naming" is operationally useful. This is standard practice in psychometrics (Messick, 1995) and educational testing but absent from agentic AI evaluation.

Provocation 2: Validate across benchmarks, not within them. Benchmark authors should coordinate a cross-benchmark evaluation study on a shared model set, subjecting score vectors to convergent-discriminant analysis (Campbell and Fiske, 1959). If scores correlate highly, the benchmarks share a common factor and one may be redundant; if they diverge, the community has identified distinct capability dimensions. Either outcome advances the field more than producing another benchmark.

On empirical validation. Reviewers rightly noted that our construct labels are interpretive rather than empirically derived. We agree, and we treat this as an open call rather than a resolved claim. The three benchmark leaderboards currently share too few models in common to yield stable rank-correlation estimates; as shared model coverage grows, a properly powered convergent-discriminant analysis (Campbell and Fiske, 1959) (requiring at minimum 20 shared models (Lord and Novick, 1968) and item-level rather than aggregate scores) becomes feasible. We intend to conduct this analysis as a follow-up and encourage benchmark authors to publish full per-model score tables to enable it.

On feasibility. We recognize that calling for a pause in benchmark production is ambitious. We do not advocate a moratorium but rather a norm: that new MCP benchmarks be accompanied by a mandatory convergent validity check against at least one existing benchmark on a shared model set. This check can be conducted post-hoc and reported in a supplementary table, adding minimal burden while substantially increasing interpretive value. Precedent exists in the NLP evaluation literature: BIG-Bench (BIG-bench authors, 2023) included cross-task correlation analysis, and HELM (Liang et al., 2023) explicitly reports inter-metric relationships.

Toward interdisciplinary consensus. The construct labels we propose (planning under ambiguity, adaptive tool mastery, operational depth) should be treated as starting hypotheses for a structured consensus process, not as the process' output. We suggest that the evaluation community convene a working group—analogous to the APA/AERA/NCME Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014)—to produce a shared ontology of agentic AI capabilities. This need not precede all future benchmarking, but it should produce at minimum a shared vocabulary and a registry of claimed constructs against which new benchmarks can position themselves.

Provocation 3: Report evaluation cost alongside accuracy. MCPMark tasks require an average of 16.2 execution turns and 17.4 tool calls per task. The computational, financial, and infrastructure costs of running these benchmarks are non-trivial

but unreported. Cost-per-evaluation-point should be a standard reporting metric, enabling practitioners to make informed decisions about benchmark feasibility (Solaiman et al., 2025).

5 Conclusion

The rapid proliferation of MCP benchmarks reflects genuine demand for evaluation infrastructure in the agentic AI ecosystem. However, benchmarks without explicit constructs are instruments without calibration: they produce numbers whose meaning remains ambiguous. We have shown that three prominent MCP benchmarks implicitly target different constructs, employ different evaluation methodologies with different reliability profiles, and face different reproducibility challenges. We have identified the conditions under which a cross-benchmark empirical validation becomes feasible and commit to pursuing it as shared leaderboard coverage grows. Our position is not that any of these benchmarks should be abandoned, but that the evaluation community should invest in construct-level infrastructure—explicit definitions, cross-benchmark validation, and a structured consensus process—before producing further task-level infrastructure.

Limitations

Our construct labels ("planning under ambiguity," "adaptive tool mastery," "operational depth") are interpretive characterizations derived from benchmark design choices, not from factor-analytic or IRT decompositions of item-level data (Embretson and Reise, 2000). The cross-benchmark empirical validation we advocate is not yet feasible given the limited overlap of models across the three leaderboards; we identify the conditions required (20+ shared models, item-level scores) and flag this as priority future work. Additionally, we focus on three benchmarks from a rapidly expanding ecosystem; a comprehensive survey would require broader coverage, including MCP-AgentBench, LiveMCP-Bench, and MCPEval.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

Anthropic. 2024. Model context protocol. https://modelcontextprotocol.io. Accessed: 2025-09-01.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Jack Block. 1995. Going beyond the five factors given: Rejoinder to costa and mccrae (1995) and goldberg and saucier (1995). *Psychological Bulletin*, 117(2):226–229.

Denny Borsboom. 2006. The attack of the psychometricians. *Psychometrika*, 71(3):425–440.

Tom Bylander. 1994. The computational complexity of propositional STRIPS planning. *Artificial Intelligence*, 69(1–2):165–204.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56(2):81–105.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Susan E Embretson and Steven P Reise. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Anuj Kumar Goyal, Peter Ku, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. In *International Conference on Learning Representations*.

Truman L Kelley. 1927. *Interpretation of Educational Measurements*. World Book Company, Yonkers-on-Hudson, NY.

Walter Kintsch. 1994. Text comprehension, memory, and learning. *American Psychologist*, 49(4):294–303.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? A meta review of evaluation failures across machine learning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Jane Loevinger. 1957. Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3):635–694.

Frederic M Lord and Melvin R Novick. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

Ziyang Luo, Zhiqi Shen, Wenzhuo Yang, Zirui Zhao, Prathyusha Jwalapuram, Amrita Saha, Doyen Sahoo, Silvio Savarese, Caiming Xiong, and Junnan Li. 2025. MCP-Universe: Benchmarking large language models with real-world model context protocol servers. *Preprint*, arXiv:2508.14704.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749.

Irene Solaiman, Zeerak Talat, and 1 others. 2025. Evaluating the social impact of generative AI systems in systems and society. In *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press.

Zhenting Wang, Qi Chang, Hemani Patel, Shashank Biju, Cheng-En Wu, Quan Liu, Aolin Ding, Alireza Rezazadeh, Ankit Shah, Yujia Bao, and Eugene Siow. 2025. MCP-Bench: Benchmarking tool-using LLM agents with complex real-world tasks via MCP servers. *arXiv preprint arXiv:2508.20453*.

Zijian Wu, Xiangyan Liu, Xinyuan Zhang, Lingjun Chen, Fanqing Meng, Lingxiao Du, Yiran Zhao, Fanshi Zhang, Yaoqi Ye, Jiawei Wang, Zirui Wang, Jinjie Ni, Yufan Yang, Arvin Xu, and Michael Qizhe Shieh. 2025. MCPMark: A benchmark for stress-testing realistic and comprehensive MCP use. *Preprint*, arXiv:2509.24002.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.