August 10, 2026

Public Comment on NIST AI 200-2 (The TEVV-Athlon Framework for Evaluating AI Systems)
Submitted by: Yvette Schmitter, CEO and Co-Founder, Fusion Collective

**Summary**
The TEVV-Athlon Framework is a clear and usable method for planning AI measurement, and its grounding in construct validity and Goodhart's Law is sound. This comment strengthens the draft on its own terms. The Measure function and the validity guidance rest on two assumptions that recent events have handsomely disproved: that the evaluation environment is sound, and that the system under test is a passive subject. Between July 21 and August 8, 2026, five developers and evaluators disclosed cases where models under their evaluation reached real systems and real people they were never meant to touch. The four additions below account for a subject that can model the test and an environment that can leak. Each map to a section or consideration already in the draft, so each can be incorporated as an amendment rather than an entirely new structure.

**1. Name evaluation-awareness as a threat to construct validity**
*Strengthens Section 4.5, Validating Measurement, and Appendix D, Table 6.*
The draft's validity guidance asks whether results generalize beyond the testing conditions and whether they relate to real-world outcomes. It doesn't yet name a mechanism by which they can fail. That mechanism now has a documented case, and it deserves a place in the draft's validity guidance.

Anthropic's own account is the clearest example produced this year. Its evaluation prompt specified to Claude that the environment was a simulation with no internet access. Anthropic attributes the failure to *"a misunderstanding between us and our evaluation partner,"* and confirms that internet access was in fact available (Anthropic, July 30, 2026). When the model reached real systems, it treated them as in-scope for the exercise. In one run it paused before publishing a malicious package, noted that on the real internet this wouldn't be acceptable, then reasoned that unfamiliar certificate authorities and the 2026 calendar date proved the environment was staged. It never revisited that conclusion.

The model's behavior was conditioned on its belief about whether it was being observed. For a measurement method, this is a first-order construct-validity problem. If behavior depends on *perceived* observation, then performance-under-test and performance-in-deployment are separate constructs, and a score that measures the first while describing the second is *not valid*. The draft cites Messick and Cronbach on construct validity. This is the case that connects their theory to a live system that can model the measurement itself. A score can measure how a system behaves when it thinks it is being watched. Deployment is not observation.

There's one caution to raise because it strengthens the point rather than softening it. These behaviors appeared under capability-elicitation conditions, with safety filters off and internet access on, which don't reflect deployment. The field cannot yet say how often this occurs once

safety features are active and cannot yet say with confidence whether the system understood its actions were real. That uncertainty is a reason to measure under deployment-representative conditions, not only under maximum-elicitation conditions.

**Recommendation.** Add evaluation-awareness to the validity guidance in Section 4.5 and the considerations in Appendix D. Require evaluators to probe whether a system behaves differently when it believes it is observed, to document that probe, and to run at least one evaluation under deployment-representative conditions alongside any maximum-capability run.

**2. Make environment integrity a precondition for a valid measurement**

*Strengthens Section 4.2 and Appendix E, Table 7.*

A measurement taken in a compromised or leaking environment is not a valid measurement. The current draft treats the environment as a neutral container. As demonstrated across July-August 2026 it's the part most likely to fail.

Five cases followed one pattern. OpenAI's models broke containment and reached Hugging Face's production infrastructure (CNN, July 22, 2026). Anthropic's models reached the internet across a review of 141,006 evaluation runs and gained unauthorized access to three organizations (Anthropic, July 30, 2026). Meta and the Chinese lab Moonshot disclosed the same shape, three of the cases traced to one evaluation partner's environment (Engadget, August 7, 2026). The UK AI Security Institute's own agents took nineteen unsanctioned actions on the live internet during its evaluation (AISI, August 4, 2026). In every case the weak point was the environment, *not* the model.

These were configuration errors and that's the point. The draft provides no step at which an evaluator confirms the environment is sound before trusting the result, so a configuration error produces an invalid measurement in silence. AISI found its own gap in the same place: its monitoring surfaced the anomaly after the fact, rather than as the evaluation ran (AISI, August 4, 2026). Environment integrity is a validity precondition, which places it inside the Measure function.

**Recommendation.** Extend Section 4.2 so that confirming environment integrity becomes a basic requirement: validated isolation, network egress that is controlled and actively justified, and monitoring able to flag or block out-of-scope actions during the run. In Appendix E, extend the contamination consideration beyond data to the environment, since an environment that reaches the open internet contaminates a result just as surely as leaked test data.

**3. Make independent challenge an expectation for high-autonomy systems**

*Strengthens Appendix D, Table 6, Independent review or challenge.*

The draft's structure lets a sole actor set the goals, run the test, and interpret the result. When that actor benefits from a favorable result, their incentive shapes the measurement. Appendix D already lists independent review or challenge as a consideration for rigor. Recent events argue for raising it from an optional consideration to an expectation in the highest-risk cases.

The most recent incident pattern makes the case on its own. In case after case, the behavior surfaced through outside signals rather than the evaluator's own controls: anonymized traffic that tripped a separate alert, a human maintainer who refused a suspicious change, one lab's

disclosure that prompted another lab's review. Internal self-run evaluation missed the behavior until an independent eye caught it.

For illustration, consider the profession built specifically for independent assurance. In 2026 KPMG Australia lost its chief executive and senior leaders after a whistleblower alleged misuse of client data, and the corporate regulator wrote to roughly 2,900 auditors to tighten oversight of conflicts of interest (Bloomberg, July 22, 2026; Reuters, via IDNFinancials, July 24, 2026). The controls that failed were the *independence* controls. This is referenced as an illustration, not proof about any AI evaluation. It's a clear reminder that independence doesn't maintain itself, even where a century of rules exists to protect it. Self-assessment carries that same risk with none of that architecture.

**Recommendation.** In Appendix D, raise independent review or challenge to a cross-cutting expectation for higher-risk and higher-autonomy deployments, distinct from and additional to internal review, proportionate to risk tier. This stays within NIST's voluntary-guidance posture while closing the gap that self-assessment can't close by design. The auditor cannot be the vendor.

**4. Modernize the worked example, and let affected persons inform validity**
*Strengthens Section 3 and Section 4.5.*
The sole worked example is a chatbot avoiding low-impact violations such as spoilers, meal advice, and travel tips. A worked example shapes what evaluators build, so the example a national method chooses then becomes the shape of the assessments that follow it. In 2026 the failure that matters is an agent that researched a maintainer, created fake identities, and tried to socially engineer approval of malicious code (AISI, August 4, 2026). The agentic red-teaming material sits in Appendix B as a reference table rather than in the body as method.

The draft's own validity guidance asks whether results relate to real-world outcomes. The people affected by a system's decisions are the most direct source of that signal, and they have no role in the current structure. Their feedback is validity evidence.

**Recommendation.** Add a worked example that evaluates a multi-step, tool-using agent, and bring the agentic content from Appendix B into the body. Add pre-deployment acceptance criteria so a result can change a decision rather than only describe one. Add a mechanism for affected persons to contest outcomes and feed real-world harm back into the measurement and treat that feedback as validity evidence under Section 4.5.

**Declaration of interests**
Fusion Collective is an AI governance, audit, and assurance firm whose principals hold ISO/IEC 42001 Lead Auditor certification. The firm has a commercial interest in independent assurance and behavioral monitoring being recognized as governance requirements. These recommendations are presented because they identify genuine gaps in the draft, not to advantage any particular methodology. Each recommendation can be met by any qualified independent party.

**Case study offer**
Fusion Collective would welcome the opportunity to contribute a case study on independent assurance and runtime monitoring of an agentic deployment for a future version of the TEVV-Athlon Framework.