CIJARA GROUP LLC
AI Advisory | Risk | Governance
Charlotte, North Carolina | CijaraGroup.com

September 16, 2026

AI Standards Team

National Institute of Standards and Technology

By email: ai-standards+tevvzd@nist.gov

RE: Input on the Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation (July 2025), submitted for subsequent iterations of the draft.

To the authors:

I am the Managing Principal of Cijara Group LLC, an AI advisory and risk governance firm in Charlotte, North Carolina. I write in my own capacity. My background is second-line model risk and operational risk at a global systemically important bank (G-SIB), including Senior Model Risk Officer and Lead Operational Risk Officer roles. On September 16, 2026 I filed a separate public comment on NIST AI 300-1 ipd, the documentation zero draft. This letter is its companion. Where the two drafts touch, I say so, because the two documents will travel to SC 42 together and should not diverge on shared terms.

The outline's note to reviewers asks for input on direction and structure, on the concept map, on specific considerations for each of testing, evaluation, verification and validation, on guidance for the Clause 5 sub-bullets, and on examples for the appendices. I offer six points in that order. Each one comes from the same experience: a validation function at a regulated institution receives evaluation results from parties it does not control and has to decide what weight to place on them. Everything below is about what a result has to carry with it to be usable by someone who was not in the room when it was produced.

Point 1. A score without its procedure is not a test result (Clause 4, Table 1)

Table 1 defines testing as a strictly specified determination according to a specified, repeatable, measurable, achievable, and relevant procedure. I support that definition and ask the zero draft to follow it to its conclusion. For a model, the procedure includes the harness or scaffold the model ran inside, the tools it had access to, the number of attempts allowed, and the effort or compute setting. Change any of those and the procedure has changed, even if the weights have not.

The effect is large. A frontier model released in September 2026 scored 62.7 percent on a public reasoning benchmark under a provider-neutral harness and 99.9 percent on the same benchmark with the same weights under a provider-supplied harness. Both figures were published. Neither publication placed the two side by side. A reader of the higher figure had no way to know the procedure differed.

Suggested text for Clause 4. A reported result qualifies as a test result only where the procedure is specified to the level needed to repeat it, including evaluation configuration. A result reported without that configuration is an evaluation finding, not a test result, and should be labeled as one. Add evaluation configuration to the Clause 4 list of what documentation should clarify, alongside the existing bullet on environmental and contextual factors pertinent to the outcome.

Point 2. Training and test overlap is a threat to validity and belongs in the framework (Clause 4 and Appendix 2)

Clause 4 promises particular focus on types of validity. Contamination, meaning overlap between the material a model was trained on and the material it is evaluated on, is the most common threat to the validity of a capability claim and the least often disclosed. The same September 2026 model scored 100 percent on a benchmark of historical vulnerabilities and 39.0 percent when the same capability was measured on vulnerabilities disclosed after its training cutoff. Both figures came from the provider. The gap is the contamination effect, and it changes the meaning of the headline number.

Suggested treatment. Name train-test overlap explicitly in the Clause 4 discussion of validity, as a validity threat rather than a data-hygiene footnote. In Appendix 2, include contamination control as a method family with at least three entries: temporal holdout, in which evaluation items postdate the training cutoff, documented decontamination of the evaluation set, and reporting a contamination-controlled result beside any headline result. Ask that reports state whether evaluation items postdate the training cutoff, and if unknown, say so.

Point 3. Reliability is carrying two meanings, and the two zero drafts are about to diverge on it (Clause 3)

Clause 3 proposes to define reliability of a TEVV approach, meaning the consistency of a measurement across repeated application. That is the right concept for a TEVV standard and matches its use in measurement science. In regulated sectors, and in NIST's own documentation zero draft, reliability is also used as a property of the evaluated system: whether the model performs consistently over repeated use in production. The documentation draft's Annex A carries robustness and reliability as distinct system properties, and my comment on that draft asked for them to stay distinct.

If the TEVV draft defines reliability one way and the documentation draft uses it the other way, the first thing an SC 42 working group will do is reconcile them, and the reconciliation will cost a revision cycle. Suggested text for Clause 3. Define measurement reliability, the consistency of a TEVV approach across repeated application, and system reliability, the consistency of the evaluated system's behavior over repeated use, as separate terms. Define robustness, the system's behavior under adverse, shifted, or out-of-distribution conditions, as a third term distinct from both. Cross-reference the documentation draft so the two use the same three labels.

Point 4. The verification-to-validation handoff is where third-party evaluation fails in practice (Clause 4, Table 1, and Clause 5)

Table 1 draws the line well. Verification checks that development outputs meet the input requirements, and it belongs to the provider. Validation checks that the product meets the requirements of the intended use, and it requires insight into the system owner and the context. In a regulated deployment those are two different organizations. The provider verifies. The deployer validates. The deployer's validation depends on evidence the provider produced during verification and does not always share.

This was the hardest category of work in second-line model validation. Third-party model providers withheld training data, equations, model logic, and code. Providers are entitled to withhold. The cost was that the withholding was undeclared, so the validating function could not scope compensating controls without negotiating the size of each gap with each provider, one at a time. The outline's Clause 5 sub-bullet on managing third parties and supply chains is the right place to address this, and the note to reviewers asks for specific guidance on each sub-bullet.

Suggested guidance for Clause 5. Name the verification-to-validation handoff as a defined point in the TEVV lifecycle. State the minimum a verifying party should make available to a validating party: the evaluation configuration under which each reported result was produced, whether evaluation items postdate the training cutoff, the conditions of any third-party evaluation, and a declaration of which items are withheld and on what basis. The declaration discloses only the fact and basis of omission, never the content. My comment on the documentation draft proposes the same declaration as a profile element under that draft's Clause 6. If both drafts carry it, a validating party has one convention to point to in a contract.

Point 5. Independence of assessors needs four disclosed attributes, not one adjective (Clause 5)

Clause 5 lists ensuring the appropriate resourcing, support, and independence of the assessors. Independence is not binary. In one recent frontier release, three external evaluators occupied three different positions on it, and a reader of the system card had to reconstruct those positions from scattered text. Four attributes determine how much weight a validating party places on a third-party evaluation, and each is a fact the parties already know.

- Commissioning relationship. Whether the provider commissioned the evaluation or the evaluator ran it at arm's length.
- Funding relationship. Who paid, and whether payment was contingent on anything.
- Evaluation duration or window. How long the evaluator had, and at what stage of the release.
- Right to publish independently. Whether the evaluator was free to publish findings without the provider's approval.

Suggested guidance for Clause 5. List these four as minimum disclosures for any third-party TEVV engagement whose results will be relied on by a party other than the commissioning organization. For Appendix 1, an example contrasting a two-week evaluation commissioned and funded by the provider, published through the provider's own release, with an arm's-length evaluation carrying an independent right to publish, would show why the four attributes matter more than the word independent.

Point 6. Structure, sector practice, and terminology

On structure, I support keeping specific methods in an appendix, aligning with ISO/IEC 42001 and ISO/IEC/IEEE 29119-1, and treating the concept map as the core contribution. A framework that survives a room of national bodies is more useful than a method catalog that is out of date at publication.

On sector practice, the note to reviewers asks how this approach fits TEVV practice from other contexts. Model risk validation in U.S. banking is a fifteen-year practice with a settled structure: conceptual soundness review, outcomes analysis, and ongoing monitoring, performed by a function independent of development. It maps onto the outline's verification and validation concepts with little translation and offers a ready source of Appendix 1 examples. One caveat belongs in the text. The current interagency guidance, Federal Reserve SR 26-2 with its OCC and FDIC counterparts of April 17, 2026, is principles-based, non-enforceable, and expressly excludes generative and agentic AI models from its scope. The practice is transferable. The authority is not.

On terminology, the outline observes that some uses of red teaming in AI conflict with other fields. I agree, and I suggest the same scrutiny for evaluation configuration, which has no settled name at all. Practitioners say harness, scaffold, setup, and conditions interchangeably. A single defined term in Clause 3 would do more for comparability than any method in Appendix 2.

Offer and disclosures

I offer to host a listening session convening model risk, vendor risk, and data governance practitioners from regulated institutions to give the TEVV team verbal feedback on the concept map and on Clause 5, consistent with the outline's invitation. I also offer to contribute one or two worked Appendix 1 examples drawn from bank model validation practice, anonymized.

Commercial interest. I advise institutions on model risk and operational risk. I hold no federal contract or subcontract related to this work and no financial interest in its outcome. If a TEVV standard of this kind is adopted, my firm could benefit from advising on its implementation. I disclose this so the input is read with that interest in view.

AI assistance. I used an AI assistant to help organize research, check citations, and draft prose. I directed the substance, verified the technical and regulatory claims against primary sources, and take responsibility for the content. The experience and judgments are my own.

I understand this input becomes part of the public record.

Respectfully submitted,

Joshua M. Reh
Managing Principal, Cijara Group LLC
Charlotte, North Carolina
Joshua@CijaraGroup.com
https://CijaraGroup.com