Scientific Methods Panel (SMP) Validity Summary

Meeting Overview

Battelle convened the Scientific Methods Panel (SMP) on March 9, 2026, to obtain expert input on challenges and emerging issues related to the validity of clinical quality measures. Measure developers and clinical stakeholders often equate "validity" with the strength of clinical evidence supporting an association between a measure's focus—typically a process or intermediate outcome—and a material health outcome. However, within Battelle's consensus-based entity (CBE) framework, such association evidence primarily supports a claim of importance, demonstrating that the measure's expected benefits outweigh potential burden or harms.

Instead, validity refers to the degree to which empirical evidence, clinical judgment, and theory justify the appropriate interpretation and use of a measure within accountability programs. Battelle requires validity evidence at both the person/encounter level and the accountable entity level to ensure that measure scores accurately and fairly represent the clinical concepts they seek to capture.

Battelle sought SMP input because measure developers sometimes misinterpret the purpose of validity evidence, submitting information that speaks to a measure's importance rather than its validity. Clarifying expectations for validity—and identifying opportunities to strengthen the quality and consistency of submissions—is essential for promoting safe and effective performance measurement.

This summary highlights the key concepts presented during the meeting, how the SMP considered each issue, and the advice and perspectives they offered to inform future guidance and decision-making.

Meeting Objectives

The objectives of the meeting were to:

- Clarify how the CBE conceptualizes validity as well as how validity differs from evidence supporting measure importance.
- Discuss common challenges and misinterpretations observed in developer submissions related to face validity, person/encounter-level validity, and accountable entity-level validity.
- Obtain SMP feedback on appropriate types of validity evidence, methodological expectations, and opportunities to strengthen guidance for developers.
- Identify areas where additional structure, clarity, or review processes may improve the accuracy and meaningfulness of measure scores.

Version 1.0 | April 2026 | *The analyses upon which this publication (or document) is based were performed under Contract Number 75FCMC23C0010, entitled, "National Consensus Development and Strategic Planning for Health Care Quality Measurement," sponsored by the Department of Health and Human Services, Centers for Medicare & Medicaid Services. Restricted:* Use, duplication, or disclosure is subject to the restrictions as stated in Contract Number 75FCMC23C0010 between the Government and Battelle.

Evidential Pluralism and Validity

Background

- Battelle described evidence and associated claims as a hierarchical system, where each successive level reflects additional data and stronger justification for validity.
  - At level 1, developers typically provide basic correlations with related process or outcome measures, with limited evidence that the correlations are grounded in theory or literature. As a result, the evidence is not sufficient to infer a causal relationship between correlated measures.
  - At level 2, developers present evidence of an association in support of a mechanism claim demonstrating that confounders are ruled out. Empirical findings support the hypothesized overlap in quality construct - grounded in literature - between the measure's focus and the comparators.
  - Levels 3 and 4 require more explicit and granular evidence and claims, which may rely on the maturity of existing evidence. Level 4 requires an experimental or quasi-experimental design with the direct manipulation of the mechanism or intervention and estimation of its effect on the measure.
- Battelle noted that many measure submissions relied on correlations alone (level 1), which are necessary but insufficient for establishing validity.

SMP Feedback

- One SMP member discussed the value of distinguishing between measures that require sophisticated validity arguments and those where the clinical concept is so straightforward that simpler evidence suffices—for example, wrong side surgery or wrong medication administration.
- Members emphasized that measures require a logic model or conceptual model that clearly describes the mechanism linking the measured process or outcome to quality. They expressed concern that current submissions often skip theoretical grounding, leaving correlations and known-groups analyses uninterpreted or conceptually unsupported.
- SMP members noted that:
  - The CBE and its corresponding processes and committee should not consider level 1 evidence (correlations) sufficient to substantiate a validity claim.
  - Developers must explicitly articulate the hypothesized mechanism and demonstrate evidence that addresses confounding and conceptual alignment.
  - Validity claims should be grounded in a theory-driven rationale, not just statistical association.
  - Higher-level evidence (levels 3-4) strengthens validity claims but may not always be feasible; what matters is clear logic, justification, and transparency.
- One SMP member stressed that validity should not become a rigid, checklist-driven exercise focused only on empirical criteria; instead, reviewers should look for whether developers clearly lay out their intended interpretation of scores and provide evidence and argumentation—empirical or conceptual—to support those interpretations.

CBE and Validity Requirements

Background

- Battelle noted that for endorsement, we ask developers to submit evidence in support of three types of validity under different contexts: face validity, person-/encounter-level (i.e., data element validity), and accountable entity-level testing. Battelle described how each type supports appropriate interpretation and use of a measure in accountability programs.

SMP Feedback

- Members discussed ongoing concerns about how measure developers select and validate codes and value sets for face validity, data element validity, and accountable entity testing. Developers sometimes use incomplete or overly broad code sets, which can cause measures to miss the core clinical concept they seek to assess.
- One member described an example in which a developer overlooked key coding guidelines and value sets, prompting the question of which type of validity should address issues related to code and value-set accuracy. Matt explained that we are considering requiring developers to attest that an independent reviewer has verified their code sets, and welcomed additional approaches to strengthen the process.
- Members agreed that developers must identify the full set of codes relevant to a clinical concept, not only the most used ones. They emphasized that incomplete or overly expansive code selections compromise measurement accuracy and the meaningfulness of a measure's score. Suggestions included engaging an impartial coding consultant to support consistent and expert value-set review.
- Members stressed that third-party review alone is not sufficient; developers also need clear coding rules and explicit justification for why each code is included or excluded. Multiple members voiced support for establishing more structured and detailed review processes for value sets to strengthen measure validity.

Face Validity

Background

- Battelle shared a recent submission in which the developer gathered feedback from technical experts and clinical staff on their level of agreement with the following items: 1) the measure is an adequate reflection of quality, and 2) the measure can distinguish good from poor quality of care. In this example, the developer did not describe how they collected feedback or whether the logic model informed the process. Further, clinicians and the technical expert panel (TEP) showed differing levels of agreement with the two statements.

SMP Feedback

- Members discussed whether measure submissions should require face validity and what constitutes sufficient evidence. They noted that current face-validity questions could be better aligned with the logic model and benefit from clearer phrasing. Members also highlighted concerns that small or highly invested groups, such as TEPs, may not provide sufficiently varied perspectives, and that broader input may yield more reliable assessments. If face validity remains a requirement, they emphasized the need for clearer guidance on appropriate participant groups and best practices for collecting and documenting evidence.
- Members generally agreed that the CBE should require face validity for both new and maintenance measures, but face validity alone is not adequate to support endorsement. They emphasized the need for ongoing reassessment of whether measure specifications and value sets accurately capture the intended clinical concepts, particularly for maintenance measures.

Person-/Encounter-Level Validity (i.e., Data Element Validity)

Background

- Battelle presented two data element validity examples.
  - The first focused on claims data, illustrating the common argument that claims are valid because they are stored in structured fields and audited.
  - The second example focused on an electronic clinical quality measure (eCQM) for which developers employed the standard approach of comparing electronically extracted data to manually abstracted data (i.e., the "gold standard").
- These examples underscore that data element validity cannot rest on assumptions about data sources or structured fields. Developers must demonstrate empirically that the data elements accurately represent the patient conditions, events, or processes the measure seeks to capture.

SMP Feedback

- Members emphasized that the accuracy of claims data varies depending on the nature of the assigned codes. For example, primary diagnosis codes are generally more accurate than secondary diagnosis codes and procedure codes. Further, reliability and validity are intertwined—a measure cannot be valid if the underlying data are not coded reliably.
- Members also raised concerns regarding the validity of claims-based diagnostic codes for complex conditions such as substance use disorder (SUD). One member indicated that patients with SUD may not have a corresponding diagnosis code, making it essential for developers to show evidence that diagnoses in claims are present, accurate, and complete in the source records. Without such evidence, between-system comparisons may be fundamentally invalid. Another member highlighted that health care providers sometimes assign codes for more than one condition during the evaluation phase of care, so allowing more than one code during a specified period may be necessary.

- Members discussed the importance of adequate sample size in validation studies that compare electronic health record (EHR) data to data from manual abstraction. Small samples can produce confidence intervals so wide that they fail to meaningfully depict data element validity. Members advised developers to clearly address sample size, precision, and confidence interval boundaries in their methodology.
- Overall, data element validity must account for differences in the nature of the assigned codes, the clinical complexity of the condition being measured, and the methodological rigor of validation studies. Measure submissions should include clear justification, transparent methodology, and empirical evidence that supports the accuracy of the specific data elements, not just assumptions about the data source.

Accountable Entity-Level Validity (i.e., Measure Score Validity)

Background

- Battelle presented two examples illustrating challenges in empiric accountable entity validity.
  - The first example was a measure composed of multiple composite measures. The developer assessed the correlation between these composites, arguing that they overlapped conceptually. However, the developer did not explain the conceptual rationale for this overlap in a hypothesis, the evidence review, or the logic model. In addition, the correlation analysis did not use an external benchmark, resulting in questions about the interpretability of the results.
  - In the second example, the developer used known-groups validity testing. The submission was not clear on whether observed differences between some groups (e.g., male vs female) reflected true differences in health care quality or simply differences in case mix; if the latter, a relevant question is whether that test is appropriate for measure score validity.

SMP Feedback

- Members observed that empirical accountable entity validity analyses often lack theoretical grounding. One member noted that face validity, while important, can be subjective, and that focusing too heavily on empirical analyses and fixed criteria without assessing conceptual clarity could be counterproductive. Measure developers should state their interpretations and provide arguments to support those interpretations, empirically or conceptually.
- One member indicated that correlation and known-groups validity differences were not satisfying in that they do not test theory-driven questions or associations. Developers should provide a logic model, or conceptual model, and an empiric evaluation of those models. However, different types of measures may require different logic or conceptual models and developers should tailor their empirical approaches accordingly.

Next Steps

Battelle will use feedback from the meeting to inform the development of a guidance document on validity for measure developers and stewards. We plan to have a draft available for member review before the June 2026 SMP meeting and to finalize the guidance document in late August 2026.