Comments on the Endorsement and Maintenance (E&M) Guidebook Version 3.0

Acumen thanks the Partnership for Quality Measurement (PQM) Consensus Based Entity (CBE) for the opportunity to provide feedback on the Endorsement and Maintenance (E&M) Guidebook Version 3.0, published in June 2025. Our comment is structured into two sections: 1) a discussion of the improvements made to the E&M process covered in the latest Guidebook, and 2) a discussion of several remaining areas for further improvements we believe PQM should consider addressing prior to finalization.

Comments

1. Improvements

We appreciate several updates made to the Guidebook which reflect feedback we previously shared with PQM. Specifically, Version 3.0 of the Guidebook:

- Adds some specificity with respect to reliability rubrics and ways for developers to address validity concerns;
- Adds an appendix section for cost measures, which:
    - acknowledges the relationship between cost and quality,
    - acknowledges the positive effects of cost measurement with respect to efficiency, value, and harm reduction,
    - includes a description of validity and notes the relevance of mechanistic studies in providing "supportive evidence of accountable entities' ability to impact the measure" (p. 72),
    - acknowledges that unintended consequences should be considered as part of Use and Usability, rather than Validity;
- Notes that space will be provided for developers to describe accountability program context under the Use and Usability evaluation rubric.

Several of these updates are direct responses to feedback our team had suggested during earlier stages of this process, so we thank PQM for their serious consideration of issues we previously highlighted.

2. Areas for Further Improvement

Despite the improvements noted above, in other instances our prior feedback about the E&M process has not been addressed. Below, we list our comments on language that is applicable to all measures, followed by comments on text specific to cost measures. Our comments encompass new Guidebook text, as well as text that has not been updated in this version of the Guidebook.

2.1. Issues Applicable to All Measures

2.1.1. *Inadequately defined and justified reliability standards*

We commend PQM's effort to provide standardized benchmarks and methodologies for determining measure reliability in the Guidebook. However, the current guidance does not adequately delineate recommended methodologies for calculating reliability or give sufficient justification for the methods it does propose. Several pieces of guidance are contradictory or unclear:

- The basis for having "not met" scientific acceptability standards is listed as having reliability scores "<0.6 for 70% or more of the accountable entities" (p. 49), but the converse criteria for having "met" reliability standards reads "[reliability] ≥0.6 for 70% or more of the accountable entities" (p. 51). We believe the intention in the "not met" case was to convey having reliability scores <0.6 for *30%* or more of accountable entities.
- While the PQM Measure Evaluation Rubric (Appendix D) lists failing reliability thresholds as grounds for having "not met" scientific acceptability standards, Table 4 seems to suggest that a measure having failed the reliability thresholds could be endorsed with proposed conditions on a 3-year maintenance cycle. We would suggest more clarity on how the reliability standards would be applied, and whether measures failing reliability thresholds could still maintain endorsement if other criteria are sufficiently met.
- The Guidebook consistently lists standards for split-half reliability based on a certain percentage of accountable entities meeting the 0.6 threshold. However, split-half reliability is not usually calculated individually for each provider given that it estimates the theoretical correlation between multiple observations of the entire set of accountable entity scores. Thus, this criterion is unclear and should be revised to give a threshold for a single intraclass correlation coefficient (ICC) value or to recommend a different reliability method (e.g., variance ratios). In the alternative, guidance should be provided as to which method(s) of calculating split-half reliability at the individual entity level are recommended.

Beyond these inconsistencies, we are concerned that there is insufficient justification given for the Guidebook's reliability standards. The updated guidance does not discuss any rationale for the specific thresholds of reliability listed (e.g., 0.6 for 70% of accountable entities), nor does it cite an external source of guidance or consensus for these exact thresholds. In addition, the Guidebook does not provide a rationale for the blanket recommendation for non-binomial measures to be assessed using split-half reliability over other methods such as signal-to-noise ratios. For binomial measures, a forthcoming paper which does not seem to be publicly available (Aume et al, 2025) is cited with no indication of how it is being referenced. The Guidebook also continues to cite the National Quality Forum's (NQF) Scientific Methods Panel (SMP) June 2022 meeting as the source for establishing reliability thresholds (p. 48). However, this meeting did not set reliability thresholds and explicitly noted that the next steps would be to hold a public comment period. This was not done.[1]

Finally, reliability standards must recognize a tradeoff with validity standards. It is straightforward to construct highly reliable measures that are clearly invalid (e.g., imagine a measure that assigns every provider's episodes a numerical score based on the first letter of the provider's name and then takes the average across episodes). More realistically, certain statistical methods may be used to enhance reliability metrics by reducing validity (e.g., shrinkage, outlier removal, increasing sample sizes by expanding inclusion criteria). It is also straightforward to construct highly valid measures that have low reliability (e.g., defining a narrow clinical cohort for a measure denominator increases the validity of most measures, while generally reducing the reliability – this effect is familiar to statisticians, as it is akin to an

[1] The reliability standards have evolved over time. There remain many unresolved aspects of SMP's years of considering reliability standards (e.g., how firmly to apply this line in recognition that any number is arbitrary, and whether additional validity testing would offset lower reliability results).

estimator that is unbiased [i.e., valid] but imprecise [i.e., less reliable], sometimes described as bias-variance trade-off). The Guidebook should address how Committee members should evaluate such tradeoffs.

2.1.2. *Lacks objective standards for validity*

The E&M process lacks objective validity standards. The Guidebook does not define explicitly validity, instead leaving its evaluation entirely up to the reviewers: the validity criterion is not met if the "[r]eviewer determines the methodology to assess validity is inadequate/inappropriate; OR the analytic approach is inadequate/inappropriate" (p. 46).

We recommend that PQM develop objective validity standards that can be applied by reviewers. We propose that the definition of validity should be the same as the Measures Management System definition for validity: "In measure development, the term "validity" has a specific application known as test validity, which refers to the degree to which evidence, clinical judgment, and theory support interpretations of a measure score. Stated more simply, test validity is an empirical demonstration of the ability of a measure to record or quantify what it purports to measure."

Furthermore, the examples provided (p. 66) with respect to correlation/causation are problematic. First, a specific correlation coefficient (0.5) is described as 'lower than expected' without any guidance explaining how such specific expectations should be established. It is not clear whether these expectations are determined based on established benchmarks/thresholds from PQM, or if the developer should pre-specify a rough expected correlation (i.e., high, medium, low, positive, negative) based on existing data, literature, or a conceptual model. Second, the suggested follow-up analysis to explore causal explanation for the correlation is more suited to researchers and/or practitioners who are able to exploit additional primary data. Measure developers that typically conduct these sort of correlation studies using generally available measures/data may face challenges beyond noting clear and well-known policy factors that could be responsible. PQM should provide a wider range of options/examples for addressing such validity concerns.

2.1.3. *Analysis requirements for Risk Adjustment*

The Guidebook includes updates to risk adjustment expectations for new versus maintenance measures. Specifically, formal analyses for risk adjustment are only required during maintenance (p. 52). However, we believe this should be more clearly stated in the description of Scientific Acceptability related to risk adjustment on page 48. PQM should also provide a rationale for why risk adjustment analyses are not required during initial endorsement. Additionally, to ensure that measures are fairly evaluated, PQM should establish that measure submissions that choose to omit optional analyses are not negatively evaluated by committee members.

2.1.4. *Inappropriately uses "acceptable model performance" as a Scientific Acceptability criterion*

We recommend removing requirements for the risk adjustment model to "demonstrate acceptable model performance" (p. 52), unless a definition of acceptable model performance is established. This requirement creates the possibility for the reviewers to be subjective and inconsistent across measures without a definition of acceptable model performance. Should

such a definition be established, we further recommend an additional public comment period so measure developers/stewards can provide input on the definition.

Currently, PQM panelists often interpret "acceptable model performance" based on explained variance metrics alone (e.g., R-squared). Such metrics are not appropriate to differentiate a good measure from a bad measure. As an illustrative example, consider an episode-based cost measure for knee replacements that deliberately does not include costs from maintenance dialysis in the numerator, because orthopedic surgeons cannot reasonably influence dialysis costs. The R-squared of a risk adjustment model with End-Stage Renal Disease (ESRD) status as a covariate will be lower in such a measure than in an alternative measure that includes maintenance dialysis in the numerator, *despite the fact that this latter alternative measure is clearly less valid*. Metrics such as predictive ratios (the ratio of observed to expected outcomes), calculated for various subsets of the measured populations (e.g., risk deciles, patients or providers with different characteristics, etc.) are much more informative about model performance and the scientific acceptability of a measure.

2.1.5. *Criteria for "Closing Care Gaps"*

We commend PQM for its attention to how quality measures can play an important role in addressing efforts to reduce gaps in care across particular sets of identifiable patient populations. As part of this effort, the criteria for maintenance states one of the requirements is the developer "*describe or provide* evidence indicating how accountable entities can utilize these results to close gaps in health care delivery and outcomes for the identified groups" (p. 43, emphasis added). However, the criteria for 'Met' only notes that there is evidence provided, rather than the more inclusive option to instead describe how providers could close care gaps using the results of the measure. Although this may be a minor distinction, we believe it to be important as this domain becomes required during the Spring 2026 cycle, in part because certain processes or outcomes may be difficult to directly measure through evidenced impact.

2.2. Issues Applicable to Cost Measures

Version 3.0 of the Guidebook adds CBE Guidance on Cost Measures. As noted above, we appreciate several aspects of this new text. However, we believe the Guidebook still falls short in terms of making meaningful and necessary improvements in how cost measures are evaluated to better support Centers for Medicare & Medicaid Services' (CMS) pursuit of high value care. Our specific comments are listed below.

2.2.1. *CBE Guidance on Cost Measures appendix*

Overall

Despite stating that the intention of this section is "is to provide information to measure developers *and E&M committee members* about the PQM's approach to reviewing cost measures," (p. 68, emphasis added), the text itself appears to be addressed primarily to developers/CMS, with no clear instructions provided to committee members. We recommend that PQM consider making guidance for developers/CMS available in other resources, if appropriate, and limiting the Guidebook to information on how PQM will consider measures for E&M.

Principles Section

While improvement in efficiency and reduction in harm are listed as primary benefits of cost measures, the sustainability of health care is not. Ultimately, the sustainability of healthcare affects access to care and, therefore, patient health outcomes and experience and should be listed as a primary benefit. The perspective of the policymakers, who are required (sometimes by law) to consider health care costs, must also be considered and emphasized.

PQM Measure Evaluation Rubric Guidance Section

- Despite stated objectives, this section does not appear to delineate any new evaluation rubrics. Instead, it lists "solutions" and "examples" for developers. This is combined with lack of meaningful change in Appendix D: PQM Measure Evaluation Rubric. Cost measure evaluation rubric guidance must include clear guidance, including met/not met criteria, for committee members.
- Reliability discussion in this section (as in the rest of the document) appears to confuse reliability with validity, and intraclass correlation with sign-to-noise reliability approaches. For example: "[Using the Intraclass Correlation Coefficient (ICC)] involves assessing the proportion of variance attributed to actual differences in care delivery *versus noise*, ensuring that the measure consistently reflects *true performance differences* across various health care entities" (p. 72, emphasis added). Care must be taken to use language appropriate to each reliability method and to remember that a statistic can measure the *wrong* thing reliably.
- Validity discussion in this section appears to advocate for risk-adjusting for endogenous treatment choices: "Explicit consideration of treatment choices, such as post-hospital institutionalization versus home health care, should be integrated into the risk adjustment covariates to enhance model performance." (p. 72) This statement must be clarified to make clear that developers are expected to not include provider actions in risk-adjustment models.
- Validity examples include mechanistic studies and risk adjustment. However, elements of the measure itself (inclusion of adverse events, exclusion of unrelated services) should be cited as acceptable ways to address validity threats.
- Usability discussion calls for specific presentation of data regarding the relationship between cost and quality (a 3-by-3 matrix showing top, mid, and bottom-ranked entities across cost and quality, as shown in Table H6-3). However, there is no information about (a) how this information will be used for evaluation (e.g., how will it supplement measure correlation), (b) whether this information is required or is optional, or (c) whether alternatives ways of presenting similar information (e.g., a scatterplot) will be acceptable. This ambiguity must be addressed.
- Row 3 of Table H6-1 should reference Table H6-3, instead of Table H6-2 (p. 69).

2.2.2. *Does not ensure that Cost and Efficiency Committee members possess a depth of familiarity with and comprehension of cost measures, healthcare efficiency, and payment policies.*

Roster categories and targets do not ensure that there are any/enough members with the needed familiarity/comprehension of cost measures. We recommend that PQM implement a nomination and selection process for Cost and Efficiency committees that ensures that expertise in cost measures and their uses are well represented.

2.2.3. *Does not describe the Scientific Methods Panel or the expertise of its members with respect to cost measures, healthcare efficiency, and payment policies.*

The Guidebook makes several references to the Scientific Methods Panel (SMP) yet provides no information about its composition and operation. PQM should follow a similarly transparent process as it uses for its other panels (e.g., hold a call for nominations, post the roster, allow the public to dial into meetings in listen-only mode, post a meeting summary or transcript). We recommend that, through this process, PQM ensure that cost expertise are sufficiently represented on the SMP.

2.2.4. *Continues to use language inappropriate for cost measures*

In multiple places, the Guidebook references "quality" instead of "performance," and/or does not mention cost (or other relevant terms). For example:

- Scientific Acceptability description: "Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the *quality of care* when implemented." (p. 46, emphasis added)
- Conceptual model instructions: "Attach a conceptual model illustrating the pathway between patient risk factors (including social, functional status-related, and clinical factors), *the quality of care*, and the measured outcome." (p. 48, emphasis added)
- "Reviewer determines methodology employed is adequate and the analytic approach presented is appropriate and thorough (e.g., clear reasoning for conducting a correlation analysis with *another* quality indicator, clear hypothesis for correlations, and supportive evidence from mechanistic studies to justify correlation results)" (p. 51, emphasis added)

The Guidebook and submission materials should use the term "performance" instead of "quality" whenever referring to general measure performance and should incorporate appropriate cost measure examples.

2.2.5. *Continues to not use cost-measure-appropriate objectives of the evaluation process.*

The Guideline states that the goal of endorsement is to ensure that measures are "safe and effective" (pp. 9, 12, 62, 68, 69, 75). The actions by measured entities in response to the measurement are what determine safety and efficacy of the act of measurement, not the measure itself. For example, a clinically valid and scientifically acceptable measure that quantifies disease detection rate may be unsafe and ineffective in a program that does not have other measures that also measure appropriate use or adverse effect of diagnostic testing. However, when the same measure is implemented alongside an appropriate use measure or a cost measure, it can be safe and effective. Therefore, safety and effectiveness cannot be evaluated in a vacuum without the policy context of how a measure will be implemented. We strongly recommend establishing that the objective of endorsement is scientific acceptability. Safety and efficacy should not be included in the E&M process, unless it is discussed in support of consideration of the scientific acceptability. For example, it may be appropriate to discuss policy/program context when considering the validity of a measure. The overall appropriateness, safety, and effectiveness of implementing a measure within a specific program should only be evaluated by the Pre-Rulemaking Measure Review (PRMR) process instead.