Measure Evaluation Report and Instructions

***INSTRUCTIONS****: This form is primarily for measure developers to use as a guide when evaluating measures. Non-CMS contracted measure developers or non-measure developers who elect to use the form for another purpose may edit the Project Overview section to reflect not having a measure development contract.

Please note: All CMS measure contract deliverables must meet accessibility standards as mandated in Section 508 of the Rehabilitation Act of 1973. This template is 508 compliant. You may not change the template format or non-italicized text. Any change could negatively impact 508 compliance and result in delays in the CMS review process. For guidance about 508 compliance, CMS’s* [*Creating Accessible Products*](https://www.cms.gov/es/node/1549751) *may be a helpful resource.

Many of the necessary elements for an expanded Measure Evaluation Report are in the Measure Information Form (MIF) and Measure Justification Form (MJF). It is important for measure developers to self-evaluate their measures iteratively throughout the Measure Lifecycle. The design of this form is to assist measure developers in comparing the specifics of their measure against evaluation criteria. The CMS MMS Hub,* [*Measure Evaluation Criteria*](https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/overview)*, describes the process in detail.

Included are specific criteria for evaluating different types of measures where they vary from or added to the general criteria. Additional instruction for evaluating the specific measure types—*[*Composite Performance Measure Evaluation Guidance*](https://mmshub.cms.gov/sites/default/files/Composite-Measures-Accountability.pdf)*,* [*Cost and Resource Use Measures*](https://mmshub.cms.gov/sites/default/files/Cost-and-Resource-Use-Measures.pdf)**, [*Electronic Clinical Quality Measures (eCQMs)*](https://mmshub.cms.gov/sites/default/files/eCQM-Specifications-Testing-Standards-Tools-Community.pdf)*, and* [*Patient-Reported Outcome Measures*](https://mmshub.cms.gov/sites/default/files/Patient-Reported-Outcome-Measures.pdf)*—are also included and noted where applicable. If there are any questions regarding a measure type, consult the* *Measures Manager staff**.****PLEASE DELETE THIS INTRODUCTORY SECTION (TEXT ABOVE THE LINE) AND REPLACE THE FORM-RELATED REFERENCES ON THE LAST PAGE OF THE FORM WITH YOUR OWN REFERENCES BEFORE SUBMISSION.*** ***CMS REQUIRES NO SPECIFIC FORMAT FOR REFERENCES BUT BE COMPLETE AND CONSISTENT.

CMS-CONTRACTED MEASURE DEVELOPERS MUST USE THE MOST CURRENT PUBLISHED VERSION OF ALL TEMPLATES AND SHOULD CHECK THE*** [***CMS MMS HUB***](https://mmshub.cms.gov/) ***FOR UPDATES BEFORE SUBMISSION.***

**Project Title: *List the project title as it should appear in the web posting.*Date:**Information included is current on *insert date*.

**Project Overview:**The Centers for Medicare & Medicaid Services (CMS) contracted with *measure developer name* to develop *measure (set) name or description*. The *contract* name is *insert name*. The *contract* number is *project number*.
 **Measure Name:

Measure Set** (or Setting):  **Measure Developer:

*Instructions:*** *For each subcriterion, enter the rating assigned using criteria from the MMS Hub sections on Measure Evaluation and the* [*CMS CBE guidance document(s)*](https://p4qm.org/EM)*, if applicable. Use the supporting information provided in the MIF and MJF, and any additional relevant studies or data. For any less-than-satisfactory ratings, enter an improvement plan in the appropriate spaces. Make a summary determination for each* [*criterion*](#Criterion) *using the subcriteria ratings with statements to support the conclusions.*

1. Evidence and Performance Gap—Importance*[[1]](#footnote-1)*

This criterion evaluates the extent to which specific measure focus is evidence-based and important to making significant gains in health care quality where there is variation in or overall, less-than-optimal performance. Measures must meet all subcriteria to pass this criterion for evaluation against the remaining criteria.

| 1. Subcriteria | Anticipated CMS Rating[H/M/L] | Rating Improvement Plan (if Low/Moderate) |
| --- | --- | --- |
| ‎1.1 Evidence to Support the Measure Focus/Measure Intent  |  |  |
| ‎1.2 Performance Gap, including disparities  |  |  |
| ‎1.3 Explicit Logic  |  |  |
| 1.4 Harmonization |  |  |

**Summary Rating for Importance:**Fail: The rating of at least one of the Importance subcriteria is not high.

Pass: Measure is important; the rating of all the subcriteria are high.

*If the measure developer plans to submit the measure to the CMS CBE for endorsement, see the CMS CBE Measure Evaluation Rubric in the* [*Endorsement & Maintenance Guidebook*](https://p4qm.org/EM)*.* ***Patient-Reported Outcome-Based Performance Measures (PRO-PM)****Patients should be involved in identifying person-centered and meaningful concepts for all quality measurement. It is* **essential** *for patient involvement in development of PRO-PMs.***Brief Statement of Conclusions Supporting the Summary Rating:

*Additional instructions and guidance for completing the table for criterion 1.***

## Evidence to Support the Measure Focus

*Health outcomes are often the preferred focus of a measure because they integrate the influence of multiple care processes and disciplines involved in the care. The measure focus is evidence-based demonstrated as*

* *Outcome—a rationale supports the relationship of the health outcome to processes or structures of care.*
* *Intermediate outcome—a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured intermediate outcome leads to a desired health outcome.*
* *Process—a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured process leads to a desired health outcome.*
* *Structure—a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured structure leads to a desired health outcome.*
* *Patient-reported evidence—should demonstrate the target population values the measured outcome, process, or structure and finds it meaningful.*
* *Efficiency— the quality component requires evidence, but not required for the resource use component. Note: Measures of efficiency combine the concepts of resource use and quality.*

*Clinical care processes typically include multiple steps:*

* *assess: identify problem or potential problem*
* *choose: plan intervention (with patient input)*
* *provide intervention*
* *Evaluate its impact on health status.*

*If the measure focus is one step in such a multi-step process, select the step with the strongest evidence for the link to the desired outcome as the focus of measurement. A measure focused only on collecting patient-reported outcome measure (PROM) data is not a PRO-PM.

The preferred systems for grading the evidence are the* [*United States Preventive Services Task Force (USPSTF)*](https://www.uspreventiveservicestaskforce.org/uspstf/grade-definitions) *grade definitions and* [*methods*](http://www.uspreventiveservicestaskforce.org/Page/Name/methods-and-processes) *or* [*Grading of Recommendation, Assessment, Development, and Evaluation (GRADE*](https://www.gradeworkinggroup.org/)***)*** *guidelines.

Present evidence for specific time frames or thresholds included in a measure. If there is a limit to the evidence, then consider the literature regarding standard norms.

Provide examples of evidence to demonstrate the target population for patient-reported measures values the measured outcome, process, or structure, finds it meaningful, and include patient input in development of the instrument, survey, or tool. Include evidence of focus group input regarding the value of the quality measure derived from the instrument/survey/tool.

Current requirements for structure and process measures (i.e., a systematic assessment and grading of the quantity, quality, and consistency of the body of evidence that the measured structure/process leads to a desired health outcome) also apply to patient-reported structure/process measures.

Target areas for PROs include health-related quality of life/functional status, symptom/symptom burden, and health-related behavior.****Instrument-Based Measures, Including PRO-PMs****Patients/persons must be involved in identifying structures, processes, or outcomes for quality measurement (i.e., person-centered, meaningful).

PRO-PMs should have the same evidence requirement as health outcomes (i.e., empirical data demonstrate the relationship of the health outcome to processes or structures of care).

Process or structure measures derived from data collected via instrument have the same evidence requirements as other structure or process measures (i.e., systematic assessment and grading of the quantity, quality, and consistency of the body of evidence linking the measured structure or process to a desired outcome).
Address exceptions to the evidence requirement for quality measures focused solely on administering a particular instrument the same way as for other measures based solely on conducting an assessment (e.g., order laboratory test, check blood pressure).* ***Cost and Resource Use Measures****For cost and resource use measures, clearly describe the intent of the resource use and the measure construct. In addition, the service categories for resource uses (i.e., types of resources or costs) included in the resource use measure are consistent with and representative of the intent of the measure.*

## Performance Gap

*It is insufficient to only state the measure is related to an important, broad topic area. Evaluate whether the measure focus is a quality problem, an opportunity for improvement with data showing considerable variation, overall, less-than-optimal performance in the quality of care across measured entities, or disparities in care across population groups. What is the anticipated impact of the measure on desired outcomes?

When assessing measure performance data for Performance Gap, the consider these factors:*

* *distribution of performance scores*
* *number and representativeness of the measured entities included in the measure performance data*
* *data on disparities*
* *size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem.*

*Examples of data on opportunity for improvement include prior studies, epidemiologic data, or data from pilot testing, or implementation of the proposed measure. If data are not available, systematically assess the measure focus (e.g., expert panel rating) and judged to be a quality problem.

Consider differently the performance gap (i.e., opportunity for improvement) for outcome measures such as mortality and patient safety events, where it may be appropriate to continue measurement even with low event rates. Process measures can reasonably reach near 100% performance with minimal opportunity for additional meaningful gains. For mortality and adverse events measures, however, it is less clear how low is attainable.

For measures using the International Classification of Diseases, 10th Revision (ICD-10) Clinical Modification (CM)/Procedure Coding System (PCS), coding gap information should be based on ICD-10-CM/PCS coded data.****Composite Measures****The composite measure as a whole must meet the performance gap criterion. Demonstrate one performance gap for each component. However, if a component measure has minimal opportunity for improvement, CMS requires justification for its inclusion in the composite (e.g., increase reliability of the composite, clinical evidence).*

***Cost and Resource Use Measures****Cost and resource use measures must demonstrate the information presented in this measurement area has a cost problem or there is variation in resources across entities.*

## Explicit Logic

### Provide a logic model (diagram) with a description of the relationships between structures and processes and the desired outcome.

*For composite measures, explicitly and logically articulate these items:*

### The quality construct, including the overall area of quality, included component measures, and relationship of the component measures to the overall composite and to each other

### The rationale for constructing a composite measure, including how the composite provides a distinctive or additive value over the component measures individually

### How the aggregation and weighting of the component measures are consistent with the stated quality construct and rationale

## Harmonization

Consider harmonization from the beginning of development of the measure. The expectation is for CMS measure developers to consider harmonization throughout the Measure Lifecycle. Either harmonize the measure specifications with related measures so that they are compatible or justify the differences.

1.4.1 Related Measures

*Harmonize specifications for this measure with related measures.*

 *or*

*Justify the differences in specifications.*

*Measure harmonization refers to the standardization of specifications for*

* *related measures with the same measure focus (e.g., influenza immunization of patients in hospitals or nursing homes)*
* *related measures with the same target population (e.g., eye exam and HbA1c for patients with diabetes)*
* *definitions applicable to many measures (e.g., age designation for children) so that they are uniform or compatible, unless justify differences (i.e., dictated by the evidence).*

*The dimensions of harmonization can include numerator, denominator, exclusion, calculation, data source, and collection instructions. The extent of harmonization depends on the relationship of the measures, evidence for the specific measure focus, and differences in data sources.*

*1.4.2 Competing Measures*

*The measure is superior to competing measures (e.g., a more valid or efficient way to measure quality).

or

Justify multiple measures.*

# [Reliability and Validity](https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/reliability)External link icon—Scientific Acceptability[[2]](#footnote-2)

Scientific acceptability is the extent to which the measure, as specified, produces consistent (i.e., reliable) and credible (i.e., valid) results about the quality of care when implemented. Measures must meet the subcriteria for both reliability and validity to pass this criterion for evaluation against the remaining criteria.

| 2. Subcriteria | Anticipated CMS Rating[H/M/L] | Rating Improvement Plan (if Low/Moderate) |
| --- | --- | --- |
| ‎2.1 Reliability |  |  |
| ‎2.1.1 Reliability Testing  |  |  |
| ‎2.2 Validity  |  |  |
| ‎2.2.1 Data Elements Correct  |  |  |
| ‎2.2.2 Exclusions  |  |  |
| ‎2.2.3 Risk Adjustment  |  |  |
| ‎2.2.4 Meaningful Differences  |  |  |
| ‎2.2.5 Comparable Results  |  |  |
| ‎2.2.6 Missing Data  |  |  |
| ‎2.3 Empirical Analysis (Composite Measures Only)  |  |  |

**Summary Rating for Scientific Acceptability:**Pass: The measure rates moderate to high on all aspects of reliability and validity.

Fail: The measure rates low for one or more aspects of reliability or validity.

**Brief Statement of Conclusions Supporting the Summary Rating:

*Additional instructions and guidance for completing the table for the scientific acceptability criterion.

Electronic clinical quality measures (eCQMs)***

*Measure developers should specify eCQMs in the Health Quality Measure Format (HQMF) and must use the Quality Data Model (QDM), Clinical Quality Language (CQL), and value sets/direct reference codes (DRCs) published on the National Library of Medicine’s (NLM’s) Value Set Authority Center (VSAC). All eCQMs must meet most evaluation criteria, same as other measures. For CMS CBE endorsement consideration, test all eCQMs for reliability and validity using the HQMF specifications.*

* *The minimum requirement is testing in* ***electronic health record (EHR) systems from more than one EHR product****. Measure developers should test on the number of EHRs they consider appropriate. It is highly desirable to test eCQMs in systems from multiple vendors.*
* *In the description of the sample used for testing, indicate how you used eCQM specifications to obtain the data.*
* *For eCQMs specified in older, previously endorsed HQMF releases, retesting is not necessary for maintenance. They may, however, need to be respecified to accommodate variations in the most current HQMF and CQL release. Test all de novo eCQMs in the most current HQMF and CQL release format.*
* *Reliance on data from structured data fields is the expectation; otherwise, show unstructured data elements as reliable. When using natural language processing to extract data elements for digital measures (including eCQMs), the measure developer should also conduct patient/encounter level reliability testing in addition to patient/encounter level validity testing.*
* *If testing of eCQMs occurs in a small number of sites, the measure developer may best accomplish it by focusing on patient/encounter-level data element validity (i.e., comparing data used in the measure to the authoritative source). However, as with other measures, test at the level of the quality measure score if the measure developer can obtain data from enough measured entities. The use of EHRs and the potential access to robust clinical data provide opportunities for other approaches to testing.*
* *If the measure developer focuses testing on validating the accuracy of electronic data, analyze the agreement between electronic data obtained using eCQM specifications and those obtained through abstraction of the entire electronic record—not just the fields used to obtain the electronic data—using statistical analyses such as sensitivity and specificity, positive predictive value, and negative predictive value. The guidance on measure testing allows this type of validity testing to also satisfy the requirement for reliability testing.*
* *Note: Testing at the level of data elements requires testing of all critical data elements—not just agreement of one final overall computation for all patients. At a minimum, assess the numerator, denominator, and exclusions (and exceptions) and report separately.*
* *CMS and the CMS CBE will not accept use of a simulated data set (e.g., Bonnie) for testing validity of data elements, but measure developers should use Bonnie for checking the eCQM specifications and logic are working as intended.*

## [Reliability](https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/reliability)External link icon

*The measure is well-defined and precisely specified for consistent implementation within and across organizations and allow for comparability.

Measure specifications include the target population (i.e., denominator) to whom the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator, target condition, event, outcome), measurement time window, exclusion, risk adjustment/stratification, definitions, data source, code lists with descriptors, sampling, and scoring/computation.

All measures using the ICD code system must use ICD-10- CM and/or PCS except for look-back periods before October 2015.****Instrument-Based Measures****Specifications for instrument-based measures also include the specific instrument (e.g., PROM); standard methods, modes, and languages of administration; whether (and how) proxy responses are allowed; standard sampling procedures; handling of missing data; and calculation of response rates to be reported with the quality measure results.****Composite Measures****Composite measure specifications include component measure specifications (unless individually endorsed); scoring rules (i.e., how the component scores are combined or aggregated); how missing data are handled (if applicable); required sample sizes (if applicable); and when appropriate, methods for standardizing scales across component scores and weighting rules (i.e., whether all component scores are given equal or differential weighting when combined into the composite).*

***Cost and Resource Use Measures****Assess cost and resource use measures on these items when evaluating the measure’s reliability*:

* *Construction logic, e.g., detail the logic steps used to cluster, group, or assign claims beyond those associated with the measure’s clinical logic.*
* *Clinical logic, e.g., detail any clustering and the assignment of codes, including grouping methodology, assignment algorithm, and relevant codes for these methodologies.*
* *Adjustments for comparability—Inclusion/exclusion criteria related to clinical exclusion, claim-line or other data quality, data validation (e.g., truncation or removal of low- or high-dollar claims, exclusion of end-stage renal disease [ESRD] patients).*
* *Adjustments for comparability—Risk adjustment - name the statistical method (e.g., logistic regression) and list all risk factor variables.*
* *Adjustments for comparability—Costing method - detail the costing method, including source of cost information; steps to capture, apply, or estimate cost information; and provide rationale for this methodology.*
* *Adjustments for comparability—Scoring, e.g., classifies interpretation of a ratio score(s) according to whether higher or lower resource use amounts are associated with a higher score, a lower score, a score falling within a defined interval, or a passing score.*

### Reliability Testing

*Reliability testing demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period, and/or the measure score is precise.

Reliability testing applies to the data elements and computed measure score. Examples of reliability testing for data elements include inter-rater/abstractor or intra-rater/abstractor studies, internal consistency for multi-item scales, and test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

Samples used for testing:*

* *Conduct testing on a sample of the measured entities (e.g., hospital, physician). The analytic unit specified for the particular measure (e.g., physician, hospital, home health agency) determines the sampling strategy for scientific acceptability testing.*
* *The sample should represent the variety of entities with measured performance. The 2010 Measure Testing Task Force recognized that samples used for reliability and validity testing often have limited generalizability because measured entities volunteer to participate. Ideally, include all types of entities with measured performance in reliability and validity testing.*
* *The sample should include adequate numbers of units of measurement and adequate numbers of patients to answer the specific reliability or validity question with the chosen statistical method.*
* *When possible, measure developers should randomly select units of measurement and patients within units.*

*For measures using ICD-10-CM/PCS coding, reliability testing should be based on ICD-10-CM/PCS coded data.*

***Instrument-Based Measures****Identify data collection instruments (e.g., tools, specific instrument, scale, single-item). Demonstrate reliability for the computed performance score. If using multiple data sources (e.g., instruments, methods, modes, languages), then demonstrate comparability or equivalency of performance scores. Specifications should include standard methods, modes, languages of administration; whether (and how) to allow proxy responses; standard sampling procedures; how to handle missing data; and calculation of response rates for reporting with the performance measure results.****Composite Measures***

*For composite measures, demonstrate reliability for the composite measure score. Testing should demonstrate that measurement error is acceptable relative to the quality signal. Examples of testing include signal-to-noise analysis, inter-unit reliability, and intraclass correlation coefficient (ICC).

Demonstration of the reliability of individual component measures is not sufficient. In some cases, component measures that are not independently reliable can contribute to reliability of the composite measure.*

## [Validity](https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/scientific-acceptability/validity)

*Evaluation of a measure’s validity involves an assessment of consistency between measure specifications and a correct, credible reflection of the quality of care provided adequately identifying differences in quality. Therefore, evaluation of a measure’s validity requires reviewing measure specifications (e.g., numerator, denominator, exclusion, risk factors) and evidence that supports them.

Measure specifications are consistent with evidence presented to support the focus of measurement. Specify the measure to capture the most inclusive target population indicated by the evidence and support exclusions with the evidence.

Measure specifications include the target population (i.e., denominator) to whom/what the measure applies, identification of those from the target population who achieved the specific measure focus (e.g., numerator, target condition, event, outcome), measurement time window, exclusion(s), risk adjustment/stratification, definitions, data sources, code lists with descriptors, sampling, and scoring/computation.*

### Data Elements Correct

*Validity testing demonstrates the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.

Validity testing applies to data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include*

* *Testing hypotheses the measure’s scores indicate quality of care (e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method)*
* *Correlation of measure scores with another valid indicator of quality for the specific topic*
* *Relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures).*

*Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process by identified experts and if specifications explicitly address whether performance scores can distinguish levels of quality. Face validity alone does not meet the criteria for a* [*fully developed measure*](https://mmshub.cms.gov/measure-lifecycle/measure-implementation/selection)*. However, face validity is acceptable for new* measures *(i.e., those not currently in use in CMS programs and undergoing substantive changes) only. P*rovide and discuss the degree of consensus and any areas of disagreement.***Composite Measures****For composite measures, empirically demonstrate validity for the composite measure score. If empirical testing is not feasible at the time of initial endorsement, acceptable alternatives include systematic assessment of content or face validity of the composite measure or demonstration that each of the component measures meet CMS CBE subcriteria for validity. By the time of endorsement maintenance, empirically demonstrate validity of the composite measure. It is unlikely a “gold standard” criterion exists, so validity testing generally will focus on construct validation—testing hypotheses based on the theory of the construct. Examples include testing the correlation with measures hypothesized to be related or not related and testing the difference in scores between groups known to differ on quality assessed by some other measure.

****eCQMs****For eCQMs, demonstrate validity at the data element level. If this is not possible, you must provide justification. For unstructured fields, the requirement is to demonstrate data element validity.****Instrument-Based Measures****Demonstrate patient/encounter-level reliability and validity for each critical new data element (i.e., instrument). Data elements previously validated do not need to undergo additional testing.****Patient-Reported Outcome Measures****For PROMs, response rates can affect validity and therefore addressed in testing. Analyze differences in individuals’ responses related to instruments or methods, modes, and languages of administration and potentially included in risk adjustment. See the* [*Risk Adjustment in Quality Measurement*](https://mmshub.cms.gov/sites/default/files/Risk-Adjustment-in-Quality-Measurement.pdf) *supplemental material for more information.*

* + 1.

### Exclusions are Supported by Clinical Evidence

*Support exclusions with clinical evidence; otherwise, exclusions are of sufficient frequency to warrant inclusion in the specification so there is distortion of the results without the exclusion.

and

If patient preference (e.g., informed decision-making) is a basis for exclusion, there must be evidence the exclusion impacts performance on the measure. In such cases, specify the measure so information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately; denominator exclusion category computed separately).

Examples of evidence an exclusion distorting measure results include*

* *frequency of occurrence*
* *variability of exclusion across measured entities*
* *sensitivity analyses with and without the exclusion.*

*Patient preference is not a clinical exception to eligibility and subject to influence by measured entity interventions.****Composite Measures****This criterion applies to the component measures and to the composite measures.*

### Risk Adjustment Strategy

*For outcome measures and other measures when indicated (e.g., resource use)*

* *when specifying an evidence-based, risk adjustment strategy (e.g., risk models, or risk stratification)*
* *is based on patient factors (including clinical and social risk factors) influencing the measured outcome and are present at start of care*
* *demonstrate adequate discrimination and calibration.*

*or

Rationale/data support no risk adjustment/stratification

Note: Do not specify risk factors influencing outcomes as exclusions. For more information, see the* [*Risk Adjustment in Quality Measurement*](https://mmshub.cms.gov/sites/default/files/Risk-Adjustment-in-Quality-Measurement.pdf) *supplemental material.

Check with your COR for the most recent guidance.****Composite Measures****Risk adjustment strategy applies to outcome component measures.*

### Meaningful Differences

*Data analysis of computed measure scores demonstrates methods for scoring and analysis of the specified measure allow for identification of statistically significant and practically/clinically meaningful differences in performance.*

*or*

*There is evidence of overall less-than-optimal performance.

With large enough sample sizes, statistically significant small differences may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point of patients who received smoking cessation counseling (e.g., 74% vs. 75%) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 vs. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across measured entities.*

### Comparable Results for Multiple Data Sources

*If specifying multiple data sources or methods, there is demonstration they produce comparable results.****Composite Measures****Comparable results for multiple data sources apply to component measures of the composite.****Cost and Resource Use****Assess cost and resource use measures on these items when evaluating the measure’s validity:*

* *adjustments for comparability—inclusion/exclusion criteria (related to clinical exclusion, claim-line or other data quality, data validation [e.g., truncation or removal of low- or high-dollar claim, exclusion of ESRD patients])*
* *adjustments for comparability—risk adjustment (name the statistical method—e.g., logistic regression— and list all the risk factor variables)*
* *significant differences in performance*
* *comparability of multiple data sources*
* *validity testing.*

### Frequency of Missing Data and Distribution

*Analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate performance results are not biased due to systematic missing data (or differences between responders and non-responders) and how the specified handling of missing data minimizes bias.*

## Empirical Support for Composite Measures

***Composite Measures****For composite measures, empirical analyses support the composite construction approach and demonstrate*

* *Component measures fit the quality construct and add value to the overall composite while achieving the related objective of parsimony to the extent possible.*
* *Aggregation and weighting rules are consistent with the quality construct and rationale while achieving the related objective of simplicity to the extent possible.*

*A composite measure must meet subcriterion 2.3 to meet the must-pass criterion of Scientific Acceptability.

If empirical analyses do not provide adequate results (or not conducted), provide other justification and If empirical analyses do not provide adequate results (or not conducted), provide other justification and acceptance received for the measure to potentially meet the must-pass criterion of Scientific Acceptability.*

*Examples of analyses*

* *if components are correlated—analyses based on shared variance (e.g., factor analysis, Cronbach’s alpha, item-total correlation, mean inter-item correlation)*
* *if components are not correlated—analyses demonstrating contribution of each component to the composite score (e.g., change in a reliability statistic), with and without the component measure; change in validity analyses with and without the component measure; magnitude of regression coefficient in multiple regression with composite score as dependent variable (**[Diamantopoulos & Winklhofer, 2001](https://journals.sagepub.com/doi/10.1509/jmkr.38.2.269.18845)**), or clinical justification (e.g., correlation of the individual component measures to a common outcome measure)*
* *ideally, sensitivity analyses of the effect of various considered aggregation and weighting rules and rationale for the selected rules; at a minimum, a discussion of the pros and cons of the considered approaches and rationale for the selected rules*
* *overall frequency of missing data and distribution across measured entities.*

*Assess composite measures as a whole in addition to the components; therefore, specifications need to include scoring, aggregation, and weighting rules. Also, assess reliability and validity for the composite rate. In some cases, components might not be independently reliable and may contribute to overall reliability of the composite measure.*

** eCQM-Specific Additional Subcriteria***There are additional or adapted subcriteria used to evaluate eCQMs:*

* *The measure is well-defined and precisely specified for consistent implementation within and across organizations, permits comparability, and has electronic measure specifications created using the HQMF specifications*
* *eCQM specifications include datatypes, categories, attributes, and entities from the QDM, value sets and/or DRCs, and measure logic from CQL*
* *Validity demonstrated by analysis of agreement between data elements exported electronically and data elements abstracted from the entire EHR and/or other electronic health data with statistical results within acceptable norms; or complete agreement between data elements and computed measure scores obtained by applying the electronic measure specifications to a simulated test data set with known values for the critical data elements*
* *Analysis of comparability of scores produced by the respecified eCQM specifications with scores produced by the original measure specifications demonstrated similarity within tolerable error limit.*
* *Provide a crosswalk of the eCQM specifications (i.e., QDM data elements, code lists, and measure logic) where a measure needs to be respecified. Note: Comparability is only an issue if maintaining two sets of specifications.*

*Measures must meet the subcriteria for both reliability and validity to pass this criterion for evaluation against the remaining criteria. If a measure does not meet all subcriteria for reliability and validity, stop; the evaluation does not proceed.*

# Feasibility*[[3]](#footnote-3)*

This criterion evaluates the extent to which required data are readily available, captured without undue burden, and implemented for performance measurement. Feasibility is important to the adoption and ultimate impact of the measure. Assess feasibility through testing or actual operational use of the measures.

| 3. Subcriteria | Anticipated CMS Rating[H/M/L] | Rating Improvement Plan (if Low/Moderate) |
| --- | --- | --- |
| ‎3.1 Data are a Byproduct of Care  |  |  |
| ‎3.2 Electronic Sources  |  |  |
| ‎3.3 Data Collection Strategy  |  |  |

**Summary Rating for Feasibility:**High/3 rating indicates that the predominant rating for most of the subcriteria is high.

Moderate/2 rating indicates that the predominant rating for most of the subcriteria is moderate.

Low/1 rating indicates that the predominant rating for most of the subcriteria is low.

**Brief Statement of Conclusions Supporting the Summary Rating:

*Additional instructions and guidance for completing the table for criterion 3.

 eCQMs****Expand the definition to the “extent to which specifications and logic require readily available data or could be captured without undue burden and implemented for performance measurement.”****Instrument-Based Measures****Minimize (e.g., availability and accessibility enhanced by multiple languages, methods, modes) the burden to respondents (i.e., people providing the data). There should be infrastructure to collect instrument-level data and integrate the collection into workflow and EHRs, as appropriate.

Minimize the burdens of data collection including those related to use of proprietary instruments and do not outweigh the benefit of performance improvement.*

## Byproduct of care (clinical measures only)

*For clinical measures, use routinely generated data used during care delivery (e.g., blood pressure, lab test, diagnosis, medication order) for the required data elements.*

## Data elements are available in EHRs or other electronic sources

*The required data elements are available in EHRs or other electronic sources. If required data are not in EHRs or existing electronic sources, specify a credible, near-term path to electronic collection.*

## Data collection strategy can be implemented

*Demonstrate how the measure developer can implement (i.e., already in operational use or testing demonstrates the strategy is ready to put into operational use) the data collection strategy (e.g., data source/availability, timing, frequency, sampling, patient-reported data, patient confidentiality).

All data collection must conform to laws regarding protected health information. Patient confidentiality is of particular concern with measures based on patient surveys and when there are small numbers of patients.**** eCQMs****While all measures require a feasibility assessment, CMS and the CMS CBE require submission of an eCQM feasibility scorecard. This feasibility assessment must address the data elements and measure logic and demonstrate the eCQM is implementable or there is adequate addressing of feasibility concerns. The feasibility assessment uses a* [*standard score card*](https://p4qm.org/sites/default/files/2023-08/eCQM-Feasibility-Scorecard.xlsx)*. Use Bonnie testing to demonstrate the measure logic will work.****Composite Measures****Criteria 3.1, 3.2, and 3.3 apply to composite measures as a whole, considering all component measures.*

# [Usability and Use](https://mmshub.cms.gov/measure-lifecycle/measure-testing/evaluation-criteria/usability)External link icon*[[4]](#footnote-4)*

Evaluation of a measure’s usability and use involves an assessment of the extent to which intended audiences (e.g., consumers, purchasers, measured entities, policy makers) could use or are using performance results for both accountability and performance improvement to achieve the goal of high-quality and efficient health care for individuals or populations.

| 4. Subcriteria | Anticipated CMS CBE Rating | Rating Improvement Plan (if Low/Moderate) |
| --- | --- | --- |
| ‎4.1 Usability  |  |
| ‎4.1.1 Improvement  |  |
| ‎4.1.2 Benefits  | [H/M/L] |
| 4.2 Use  |  |  |
| ‎4.2.1 Accountability and Transparency  |  |  |
| ‎4.2.2 Measure Feedback  | [Pass/No Pass] |  |

**Summary Rating for Usability:**High rating indicates that the predominant rating for most of the subcriteria is high.

Moderate rating indicates that the predominant rating for most of the subcriteria is moderate.

Low rating indicates that the predominant rating for most of the subcriteria is low.

**Brief Statement of Conclusions Supporting the Summary Rating:

*Additional instructions and guidance for completing the table for criterion 4.****CMS and the CMS CBE may consider important outcome measures without an identified improvement because there is an expectation of usefulness by informing quality improvement. They inform quality improvement by identifying the need for stimulating new approaches to improvement.****Composite Measures****CMS CBE endorsement applies only to the usability of the composite measure as a whole, not to the individual component measures—unless submitted and evaluated for individual endorsement.****Instrument-Based Measures****Provide adequate demonstration of the criteria supporting usability and ultimately use of an instrument-based measure for accountability and performance improvement.

An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement.*

## Usability

### Improvement

*Demonstrate progress toward achieving the goal of high-quality, efficient health care for individuals or populations. If not in use for performance improvement at the time of initial endorsement, then a credible rationale describes how to use the performance results to further the goal of high-quality, efficient health care for individuals or populations.

An important outcome that may not have an identified improvement strategy still can be useful for informing quality improvement by identifying the need for and stimulating new approaches to improvement. Demonstrated progress toward achieving the goal of high-quality, efficient health care includes evidence of improved performance and/or increased numbers of individuals receiving high-quality health care. Consider exceptions with appropriate explanation and justification.

Demonstrate progress of the quality measure in facilitating progress toward achieving the goal of high-quality, efficient health care for individuals or populations outweighs evidence of unintended negative consequences to individuals or populations (if such evidence exists).****Composite Measures****Improvement applies to composite measures.*

### Benefits

*Benefits of the quality measure in facilitating progress toward achieving high-quality efficient health care outweigh the evidence of negative unintended consequences to individuals or populations (if such evidence exists).****Composite Measures****Benefits apply to composite measures and component measures. If there is evidence of unintended negative consequences for any of the components, the measure developer should explain how to handle or justify why that component should remain in the composite.*

## Use

### Accountability and Transparency

*Use performance results in at least one accountability application within 3 years after initial endorsement and publicly reported within 6 years after initial endorsement (or the data on performance results are available). If not in use at the time of initial endorsement, then provide a credible plan for implementation within the specified time frames.

Transparency is the extent to which performance results are identifiable, there is disclosure of accountable entities, and available outside of the measured entities with measured performance. Achieve maximal transparency with public reporting, defined as making comparative performance results about identifiable, measured entities freely available (or at nominal cost) to the public (i.e., generally on a public website). At a minimum, the data on performance results about identifiable, measured entities are available to the public (e.g., in an unformatted database). The capability to verify the performance results adds substantially to transparency.

Accountability applications are uses of performance results about identifiable, measured entities to make judgments and decisions as a consequence of performance, such as reward, recognition, punishment, payment, or selection (e.g., public reporting, accreditation, licensure, professional certification, health information technology incentives, performance-based payment, network inclusion/ exclusion). Selection is the use of performance results to make or affirm choices regarding measured entities of health care or health plans. Note: A credible plan includes the specific program, purpose, intended audience, and timeline for implementing the measure within the specified time frames. A plan for accountability applications addresses mechanisms for data aggregation and reporting.

CMS considers measures included on the Care Compare or similar website as publicly reported.*

### Measure Feedback

*Demonstration of the feedback on the measure by the measured entities or by others is when*

* *The measured entities receive performance results or data, as well as assistance with interpreting the measure results and data*
* *The measured entities and other users receive an opportunity to provide feedback on the measure performance or implementation*
* *Consideration of this feedback occurs when the measure incorporates the changes into the measure.*

*There is no intent to interpret this guidance as favoring measures developed by organizations able to implement their own measures (e.g., government agencies, accrediting organizations) over equally strong measures developed by organizations that may not be able to do so (e.g., researchers, consultants, academics). Measure developers may request a longer time frame with appropriate explanation and justification.* ***Composite Measures****Measure feedback applies to composite measures. To facilitate transparency, at a minimum, list the individual component measures of the composite with use of the composite measure.*

**Preliminary Recommendation for Adoption***Based on the individual rating of each of the four major criteria, provide an initial recommendation for adoption based on the overall suitability of this measure by marking an X in the appropriate boxes.*

| **Criteria** | High | Medium | Low | Insufficient |
| --- | --- | --- | --- | --- |
| 1. Importance |  |  |  |  |
| 2.1 Overall Reliability |  |  |  |  |
| 2.2 Overall Validity |  |  |  |  |
| 3. Feasibility |  |  |  |  |
| 4. Usability and Use |  |  |  |  |

**Recommendation:

Explanation:**

*[Please delete this list of references and replace them with your own references before submission]*

References

Centers for Medicare & Medicaid Services. (n.d.). *Creating accessible products*. Retrieved October 30, 2023, from <https://www.cms.gov/es/node/1549751>

Diamantopoulos, A., & Winklhofer, H.M. (2001, May). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*, No. 2, pp. 269-277. [https://doi.org/10.1509/jmkr.38.2.269.18845](https://doi.org/10.1509/jmkr.38.2.269.18845%20)

Grading of Recommendations Assessment, Development and Evaluation Working Group. (n.d.). *GRADE*. Retrieved October 30, 2023, from <https://www.gradeworkinggroup.org/>

Partnership for Quality Measurement. (n.d.). Endorsement and maintenance (E&M). Retrieved September 28, 2023, from <https://p4qm.org/EM>

U.S. Preventive Services Task Force. (n.d.). *Grade definitions*. Retrieved October 30, 2023, from <https://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions>

1. The need to provide an improvement plan versus a rationale should be determined by the type of measure. For process measures, a low or moderate rating requires an improvement plan. For outcome measures, a low rating requires an improvement plan. A moderate rating requires only a rationale. For CMS-contracted measure developers, check with your Contracting Officer’s Representative (COR) for specifics. [↑](#footnote-ref-1)
2. The need to provide an improvement plan versus a rationale should be determined by the type of measure. For process measures, a low or moderate rating requires an improvement plan. For outcome measures, a low rating requires an improvement plan. A moderate rating requires only a rationale. For CMS-contracted measure developers check with your COR for specifics. [↑](#footnote-ref-2)
3. The need to provide an improvement plan versus a rationale should be determined by the type of measure. For process measures, a low or moderate rating requires an improvement plan. For outcome measures, a low rating requires an improvement plan. A moderate rating requires only a rationale. For CMS-contracted measure developers, check with your COR for specifics. [↑](#footnote-ref-3)
4. The need to provide an improvement plan versus a rationale should be determined by the type of measure. For process measures, a low or moderate rating requires an improvement plan. For outcome measures, a low rating requires an improvement plan. A moderate rating requires only a rationale. For CMS-contracted measure developers, check with your COR for specifics. [↑](#footnote-ref-4)