Composite Measures for Accountability Programs

This document provides information about composite measureⓘ specificationsⓘ and discusses development, measure testingⓘ, and evaluation of composite measures intended for quality measurement in accountability programs. The Blueprint content does not cover quality indicator aggregations such as the Nursing Home Compare star rating and other similar collections of quality measuresⓘ, and instead focuses on the development, implementation, and maintenanceⓘ of individual quality measures. Composite measures can be useful for pay-for-performance programs and public reporting websites because they take several components and combine them into a single metric summarizing overall performance. This information supplements the content found in the Blueprint content on the *CMS MMS Hub*, *Measure Specification*⧉ and *Measure Testing*⧉.

1. COMPOSITE MEASURE DEFINITION

CMS defines a composite measure as "a measure that contains two or more individual measures, resulting in a single measure and a single score." There are two primary types of composite measures:

- Measures of two or more individual performance areas scored using an algorithm producing a single score as its only output. With this type of composite, the individual components cannot produce individual scores, e.g., the measured entity meets all of the composite components or none of the components.

- Measures with two or more individual component measures assessed separately and then aggregated into one score. Component elements of this type of composite stand alone, but their combination produces a richer representation of the target construct (e.g., Global Malnutrition Composite (CMIT Measure ID 1611 ⬀).

Other names for composite measures are composite performance measure, composite index, composite indicator, summary score, summary index, or scale. Composite measures can evaluate various levels of the health care system such as individual patient data, individual practitioners, practice groups, hospitals, or health care plans.

2.   PURPOSE OF COMPOSITE MEASURES

When measure developers group measures as a composite, they must have a purpose for the use of the composite (e.g., comprehensive assessment of adult cardiac surgery quality of care). There also needs to be a delineated quality construct for measurement (e.g., the four domains of cardiac surgery quality, which include perioperative medical care, operative care, operative mortalityⓘ, and postoperative morbidityⓘ).

Measure development is unique for composites because the measure developer should examine and understand the intended use of the composite and relationships between the component measures. The American College of Cardiology Foundation/American Heart Association Task Force on Performance Measures (2010) (PDF) ⬀ provides useful guidance for composite measure development:

- Explicitly state the purpose, intended audience, and scope of a composite measure.
- The individual measures comprising a composite measure should be evidence-based, valid, feasible, and reliable.
- The methods for weighting and combining individual measures into a composite measure should be transparent and empirically tested.
- Demonstrate the scientific properties of these measures, including reliabilityⓘ, accuracy, and predictive validityⓘ.
- Composites should be useful for clinicians and/or payers to identify areas for quality improvement.

The American College of Cardiology Foundation/American Heart Association Task Force on Performance Measure implements this guidance when developing new measures and updating existing measures (American College of Cardiology Foundation/American Heart Association Task Force on Performance Measure, 2019 ⬀; American College of Cardiology Foundation/American Heart Association Task Force on Performance Measure, 2020 ⬀).

3.   COMPONENT MEASURES

A composite may comprise specified and endorsed component measures; however, this is not a CMS or a CMS consensus-based entity (CBE) requirement. The *Composite Performance Measure Evaluation Guidance* ⬀ provides direction to measure developers who are selecting measures for inclusion in a composite:

- Justify the components based on evidence.
- Justify measures in terms of feasibilityⓘ, reliability, and validityⓘ.

- Individual components generally should demonstrate a gap in care; however, if included, make a clinical or analytic justification for including components that do not demonstrate a gap in care.
- Individual components may not be sufficiently reliable independently, but include them if they contribute to the reliability of the composite.

Measure developers should assess the components of the composite for internal consistency. Internal consistency is the extent to which several measures of a given construct provide similar information about that construct. Although no longer an active measure in a CMS program, Optimal Diabetes Care (CMIT Measure ID 1516⊠)(CMS CBE 0729), provides a good example. The consensus endorsement entity agreed with the measure steward ⓘ the optimal management of hemoglobin A1c, blood pressure, statin use, tobacco non-use, and daily aspirin or anti-platelet use for patients with diagnosis of ischemic vascular disease adequately represented excellent management of diabetes mellitus by preventing or reducing future complications associated with poorly managed diabetes. Each of these measures individually represent good care of diabetes symptoms, and as a group are internally consistent with the construct of comprehensive diabetes management. Consistency may be less relevant if the goal of the composite is to combine multiple distinct dimensions of quality rather than a single dimension. Standard psychometric criteria would not apply to that scenario; therefore, it may be difficult to evaluate internal consistency for composites with multiple distinct dimensions.

4.   COMPOSITE MEASURE SPECIFICATIONS

Although the measure developer may have documented the technical specifications of all components of the composite previously, they must complete the specifications for the composite. The composite measure as a whole must meet evaluation criteria; however, the component measures may not meet all the evaluation criteria. Descriptions of the criteria for testing and evaluating composite measures are in sections 6 and 7 of this document.

The methodology and considerations for weighting and scoring ⓘ include ensuring the weighting and scoring of components support the goal articulated for the composite measure. Then, using a specified method, the measure developer combines the component scores into one composite. Newer composite measures may use machine learning to weight and score component measures.

5.   COMMON TYPES OF COMPOSITE MEASURES

Table 1 provides descriptions of five common types of composite measure scoring. This list is not exhaustive; there is allowance for other scoring methods. Table 1 includes some advantages and disadvantages for each type with examples of measures in the category. The five types discussed are

- all-or-none
- any-or-none
- linear combinations
- regression-based composite measures
- opportunity scoring

Table 1. Types of Composite Measure Scoring

Type of Scoring	Advantages	Disadvantages	Examples/Evidence
All-or-None (Defect-free Scoring**)** Process Measures The patient is the unit of analysis. Only count as successes those patients who received all indicated processes of care. For all-or-none scoring, the Blueprint defines performance as the proportion of patients receiving all specified care processes for which they were eligible. No credit given for patients who receive some, but not all required items.	• Promotes a high standard of excellence. • Patient-centric. • Fosters a systems perspective. • Offers a more sensitive scale for assessing improvements. • Especially useful for those conditions for which achieving a desired clinical outcome empirically requires reliable completion of a full set of tasks (i.e., when partial completion does not gain partial benefit).	• May waste valuable information since the measure may ignore some successes. • May inadvertently weight common, but less important processes more heavily than infrequent, but important processes. • The measured entity who achieved four of five measures appears the same as the measured entity who achieved none of five measures. • The all-or-none approach will amplify errors of measurement (e.g., one unreliable component measure will contaminate the whole score), so it is essential that each of the component measures be well designed.	• Minnesota Community Measurement Optimal Diabetes Care measure. • IHI Bundles: ventilator, central line. • Society of Thoracic Surgeons (STS) Perioperative Medical Care, a process bundle of four medications: preoperative beta blockade and discharge anti-platelet, beta blockade, and lipid-lowering agents. • Study using Premier Surgical Care Improvement Project (SCIP) (Stulberg et al., 2010 ⧉) data; adherence measured through a global all-or-none composite infection-prevention score was associated with a lower probability of developing a postoperative infection. However, adherence reported on individual SCIP measures was not associated with a significantly lower probability of infection.
Any-or-None Process or Outcome Measures Similar to all-or-none, but used for events that should not occur. The patient is the unit of analysis. Any-or-none counts a patient as failing if they experience at least one adverse outcome from a list of two or more adverse outcomes.	• Promotes a high standard of excellence. • Useful when component measures are rare events.	• Particularly problematic when mixing rare, but important outcomes with common but relatively unimportant outcomes because the outcome that occurs most frequently is likely to dominate the composite.	• STS Postoperative Risk-Adjusted Major Morbidity, which is any of the following: renal failure, deep sternal wound infection, re-exploration, stroke, and prolonged ventilation/ intubation. This is an "any-or-none" measure requiring the absence of all such complications.

Type of Scoring	Advantages	Disadvantages	Examples/Evidence
Linear Combinations Can be simple average or weighted average of individual measure scores.	• Simplicity • Transparency • Linear combinations are best when supported by a strong conceptual rationale. Two frequently cited rationales are competing or uncertain importance of the component measures. An example of a competing importance rationale is a composite that includes both mortality and readmissions components (i.e., improving one may or may not improve the other). An example of an uncertain importance rationale is a composite that includes components that may or may not be relevant to a particular user.	• Does not account for potential differences in the validity, reliability, and importance of the different individual measures (Peterson et al., 2010⬈). • Equal weighting may be undesirable if there is a considerable imbalance in the numbers of measures from different domains. • Different interested parties have different priorities; one weighting method may not meet the needs of all potential users (Peterson et al., 2010⬈). • When averaging items with a small standard deviation with items with a large deviation, items with the large standard deviation tend to dominate the average. • If combining items that are not positively or negatively correlated with one another (i.e., co-vary), the resulting composite score may not possess reasonable properties to enable meaningful differentiation among patients and may not measure a single construct. Measure developers can mitigate the issue by pursuing latent factor analysis strategies to ensure that items cohere to form a reasonable single score for a construct.	• The Premier Hospital Quality Incentive (HQI) Demonstration used a composite of process and outcome measures to measure quality for coronary artery bypass graft (CABG). The composite quality score (CQS) was based on an equally weighted combination of seven measures (i.e., four process measures and three outcome measures). The publicly reported data suggest the process measures more heavily influenced the CQS than expected by the apparent 4:3 weighting. • The U.S. News & World Report Index of Hospital Quality for cardiology and heart surgery is a linear combination of three equally weighted components: reputation, risk-adjusted mortality, and structure. Although the Index weights the three components equally, a hospital's reputation score has the highest correlation with its overall score. In comparison, the Mortality Index appears to have much less influence. • The Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSI) composite measure (i.e., PSI 90) uses a weighted average of various individual component measures. The weighting was determined by an expert panel.

Type of Scoring	Advantages	Disadvantages	Examples/Evidence
Regression-based Composite Measures If the gold standard is a certain outcome, the weighting of individual items may be determined empirically by optimizing predictability of the combined items in matching the gold standard end point.	• The weight assigned to each item is directly related to its reliability and the strength of its association with the gold standard end point. • Regression-based weighting may be appropriate for predicting specific end points of interest.	• Weighting may not be optimal for objectives such as motivating health care professionals to adhere to specific treatment guidelines.	• The Leapfrog Group developed surgical "survival predictor" composite measures to forecast hospital performance based on prior hospital volumes and prior mortality rates. They used an empirical Bayesian approach to combine mortality rates with information on hospital volume at each hospital. The measure weights the observed mortality rate according to how reliably it is estimated, with the remaining weight placed on hospital volume.
Opportunity Scoring Opportunity scoring counts the number of times a measured entity performs a given care process (numerator) divided by the number of chances a measured entity had to give this care correctly (denominator). Unlike simple averaging, this method implicitly applies weighting to each item in proportion to the percentage of eligible patients, which may vary from measured entity to measured entity.	• Provides an alternative to simple averaging often used for aggregating individual process measures. • Increases the number of observations per unit of measurement, potentially increasing the stability of a composite estimate, particularly when the sample size for individual measures is not adequate.	• The most common care processes influence rate, regardless of whether they are the most important methods.	• The Hospital Core Performance Measurement Project for the Rhode Island Public Reporting Program for Health Care Services developed the opportunity model in 1998. • The Premier HQI Demonstration used the opportunity scoring method for the process composite rate for each of five clinical areas. Divide the sum of all numerators by the sum of all denominators in each clinical area.

⌨The *Measure Authoring Tool* (MAT)⧉ currently supports electronic clinical quality measure (eCQM)ⓘ composite measures within the metadataⓘ section. The MAT limits measure scoring to All-or-Nothing, Opportunity, or Patient-level Linear. Currently, users who are defining a composite measure can indicate the measure as composite and can then identify measures in the metadata that are component measures. Find more information in the MAT User Guide⧉.

6.   Composite Measure Testing

The use of composite measures creates unique issues associated with measure testing. See the Blueprint content on the *CMS MMS Hub*, *Measure Testing* ⤢ for more information.

6.1   Component and Composite Reliability and Validity Testing

Scientific acceptabilityⓘ for composite measures needs to demonstrate the component measures and the composite measure are reliable and valid. Measure developers should treat each component measure as a standalone measure, i.e., each measure will go through all the stages of the Measure Lifecycle.

The recommendation is to demonstrate reliability and validity for the composite and the components of the composite. However, demonstration of the reliability of the individual components is insufficient. It is possible for individual components to contribute to the reliability of the composite without being independently reliable. For the composite scoreⓘ, the measure developer must demonstrate the validity empirically. Much like validity testingⓘ for single measures, validity testing for the composite should also include reporting of the overall frequency of missing data and distribution across measured entitiesⓘ. It is ideal to report the effect of alternative rules for handling missing data and the rationale for the selected approach. The measure developer will discuss the pros and cons of the approaches and the rationale for the selected rules. If submitting the composite for endorsement, check the CMS CBE webpage ⤢ as they may have additional requirements.

6.2   Component Coherence

Measure developers should test to determine whether components of a composite measure adequately support the goals articulated in the constructs for the measure. In addition, measure developers should test reliability of the components using correlation analyses or confirmatory factor analysis methods. If components are coherent, the component items meet the intent of the measure construct.

6.3   Composite-Specific Testing

Components of a composite measure should support the overall goal of the measure. If components are correlated, testing analysis should be based on shared variance such as factor analysis, Cronbach's alpha, item-total correlation, and mean inter-item correlation. If components are not correlated, testing should demonstrate the contribution of each component to the composite score.

For example
- a change in a reliability statistic such as intra-class correlation coefficientⓘ, with and without the component measure
- a change in validity analyses, with and without the component measure
- the magnitude of a regression coefficient in multiple regression with a composite score as a dependent variable
- the clinical justification demonstrating correlation of the individual component measures to a common outcome measureⓘ

6.4   Appropriateness of Aggregation Methods

When aggregating components for a composite measure to explain an outcome, measure developers should identify the method they used to estimate the composite score and test the validity of the score.

When scored, the measure developer should present the results with justification of the methods used to estimate the composite score because the method selected for combining components may influence interpretation of a composite measure result.

6.4.1   Selecting Appropriate Method to Test for Composite Validity

Testing should include an examination of the appropriateness of the method(s) the measure developer used to combine the components into an aggregate composite score. For example, the testing (i.e., assessment) of a weighting methodology for process measuresⓘ may include examining the adequacy of all-or-none, any-or-none, if/then, or opportunity scoring approaches used to create the composite. For a composite outcome that uses differential weighting of the components, the documented support for the weighting methodology might include a regression of a gold standard outcome upon the components. When using a linear combination to create a composite, the measure developer should assess the components of the composite for their contribution to the validity of the overall composite score. Linear combination alone does not imply equal or differential weighting or the appropriateness of retained components within a composite score.

6.4.2   Justification of Methodology Selected

Regardless of whether the combination of the components is with equal or unequal weighting, the composite development methodology needs to include a justification for the inclusion or retention of each contributing component in the composite. Measure developers should provide specific explanations for the decisions surrounding both weighting and component retention. In addition, assessment methods should include a description of how the composite's components relate to one another regarding the decisions on component retention and weighting.

If most of the composite's variation is the result of only a subset of the components comprising the composite, the measure developer should also provide information (e.g., a table) on the contribution of each of the components to the composite (e.g., regression coefficients or factor loadings) to address which subset of components is contributing to the majority of the aggregate's variation. The measure developer may convey the variation (i.e., information content) of a composite in a variety of ways, such as through reporting of regression results, factor loadings, and percentages of shared variation explained from a principal components analysis.

Alignment of the results of the composite evaluation process might not be with the separate results for each of the components in the composite measure, as the composite may primarily reflect a minority of the components of the composite. For example, group differences on an emergency department (ED) composite measure may be largely determined by ED wait times because variability for this component may be large relative to the variability of all remaining composite components. The measure developer may resolve this issue by providing tables showing the weights or loading for each composite such that a reader can determine the impact of differential weighting on the meaning of the overall composite measure.

Measure developers should provide information for variable or component-within-composite retention decisions. For example, when using a stepwise regression model, one often selects the default values for entering and removing variables (i.e., for entry, $p < 0.05$; for removal, $p < 0.10$). When using composites created through principal component analysis or other factor analytic models, a table should show the item loadings (i.e., a type of weighting) and contain a note if there was use of other inclusion or exclusion criteria.

Measure developers should also assess the appropriateness of methods to address component missing data when creating the composite score. This analysis of missing component scores should support the specifications for scoring and handling missing component scores.

Examples of resources for methodology include[1]

- Schwartz, M., Restuccia, J. D., and Rosen, A. K. (2015). Composite measures of health care provider performance: A description of approaches. *The Milbank Quarterly*, *93*(4), 788–825. https://doi.org/10.1111/1468-0009.12165⬈
- Shahian, D. M., He, X., Jacobs, J. P, Kurlansky, P. A., Badhwar, V., Cleveland Jr., J. C., Fazzalari, F. L., Filardo, G., et al. (2015). The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: A report of the Society of Thoracic Surgeons quality measure task force. *The Annals of Thoracic Surgery, 10*(4), 1315-1325. https://doi.org/10.1016/j.athoracsur.2015.06.122⬈
- National Quality Forum. *(2021, August 30). Developing and testing risk adjustment models for social and functional status-related risk within healthcare performance measurement.* https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=96087⬈

6.4.3   Feasibility and Usability of Composite Components

Measure testing may also demonstrate that measured entities can consistently implement the measure across organizations by quantifying comparable variation for individual components, the group/organization level can deconstruct the measure into its components to facilitate transparency, and the intended measure audience can understand the measure.

7.   Composite Measure Evaluation

There are unique issues associated with composite approach that require additional evaluation. The validity of the component measures, the appropriateness of the methods for scoring/aggregating and weighting the components, and interpretation of the composite score all require evaluation. The measure developer should evaluate both the composite and its component measures to determine the suitability of the composite measure. When evaluating composite measures, measure developers should use the measure evaluation criteria, subcriteria, and special considerations. Information from the *Composite Performance Measure Evaluation Guidance*⬈ describes an approach to evaluation.

A coherent quality construct and rationale for the composite measure are essential for determining

- which components to include in a composite measure
- aggregation and weighting of the components
- which analyses to use to support components and demonstrate reliability and validity
- added value over that of individual measures alone

Reliability and validity of the individual components do not guarantee reliability and validity of the constructed composite measure. The measure developer should demonstrate the reliability and validity of the constructed composite measure while considering these items.

- When evaluating composite measures, consider both the quality construct itself and the empirical evidenceⓘ for the composite (i.e., supporting the method of construction and methods of analysis).

---

[1] This is not an exhaustive list, nor is it making a judgment that these sources are the best sources.

- Each component of a composite measure should provide added value to the composite as a whole—either empirically (because it contributes to the validity or reliability of the overall score) or conceptually (for evidence-based theoretical reasons). Choose the smallest set of component measures possible. However, including measures from all necessary performance domains may be conceptually preferable to eliminating measures because they do not contribute as much statistically.
- Individual components in a composite measure may or may not be correlated, depending on the quality construct.
- Aggregation and weighting rules for constructing composite measures should be consistent with the quality construct and rationale for the composite. A related objective is methodological simplicity. However, complex aggregation and weighting rules may improve the reliability and validity of a composite measure, relative to simpler aggregation and weighting rules.
- Standard CMS CBE measure evaluation criteria apply to composite measures.
- Note: The CMS CBE only endorses composite measures intended for use in both performance improvement and accountability applications.

8. KEY POINTS

Composite measures combine two or more component measures, each of which individually reflects quality of care, into a single quality measure with a single score. These measures can be useful in pay-for-performance programs and public reporting websites because they take several components and combine them into a single metric summarizing overall performance. There are several different types of composite measures, including all-or-none, any-or-none, linear combinations, regression-based composite measures, and opportunity scoring. Each of these composite measure types has unique advantages and disadvantages that measure developers should consider when determining if and how to develop a composite measure.

Composite measures undergo the same processes for development and testing as other measures, however they have some additional requirements: measure developers must conduct scientific acceptability and feasibility testing for both the full composite measure as well as for each of the individual components. This testing includes evaluation of the aggregation methods (i.e., the methods used to combine the components into a total score). Regardless of the aggregation method, the measure developer must provide justification for the inclusion or retention of each contributing component, along with justification for the component weighting in the total composite score.

REFERENCES

Casey Jr, D. E., Thomas, R. J., Bhalla, V., Commodore-Mensah, Y., Heidenreich, P. A., Kolte, D., Muntner, P., Smith Jr, S. C., Spertus, J. A., Windle, J. R., Wozniak, G. D., Ziaeian, B. (2019). 2019 AHA/ACC clinical performance and quality measures for adults with high blood pressure: A report of the American College of Cardiology/American Heart Association Task Force on Performance Measures. *Circulation: Cardiovascular Quality and Outcomes 12*, 11. https://doi.org/10.1161/HCQ.0000000000000057

Centers for Medicare & Medicaid Services. (n.d.-a). *Measure authoring tool*. Retrieved November 29, 2022, from https://www.emeasuretool.cms.gov/

Centers for Medicare & Medicaid Services. (n.d.-b). *Global malnutrition composite score*. Retrieved July 19, 2023, from https://cmit.cms.gov/cmit/#/FamilyView?familyId=1611

Centers for Medicare & Medicaid Services. (n.d.-c). *Optimal diabetes care*. CMS Measures Inventory Tool. Retrieved November 29, 2022, from https://cmit.cms.gov/cmit/#/FamilyView?familyId=198

Heidenreich, P. A., Fonarow, G. C., Breathett, K., Jurgens, K. Y., Pisani, B. A., Pozehl, B. J., Spertus, J. A., Taylor, K. G., Thibodeau, J. T., Yancy, C. W., Ziaeian, B. (2020). 2020 ACC/AHA clinical performance and quality measures for adults with heart failure. *Journal of the American College of Cardiology 76*, 21. https://www.jacc.org/doi/10.1016/j.jacc.2020.07.023

National Quality Forum. (2013). *Composite performance measure evaluation guidance.* https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=73046

National Quality Forum. (2021, August 30). *Developing and testing risk adjustment models for social and functional status-related risk within healthcare performance measurement.* https://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=96087

Partnership for Quality Measurement. (n.d.). *Endorsement and Maintenance*. Retrieved July 19, 2023, from https://p4qm.org/EM

Peterson, E.D., DeLong, E.R., Masoudi, F.A., O'Brien, S.M., Peterson, P.N., Rumsfeld, J.S, Shahian, D. M., & Shaw R.E. (2010). ACCF/AHA 2010 position statement on composite measures for healthcare performance assessment: a report of the American College of Cardiology Foundation/American Heart Association task force on performance measures (writing committee to develop a position statement on composite measures). *Circulation*, *121*, 1780 –1791. https://doi.org/10.1161/CIR.0b013e3181d2ab98

Schwartz, M. Restuccia, J. D., and Rosen, A. K. (2015). Composite measures of health care provider performance: a description of approaches. *The Milbank Quarterly, 93*(4), 788–825. https://doi.org/10.1111/1468-0009.12165

Shahian, D. M., He, X., Jacobs, J. P, Kurlansky, P. A., Badhwar, V., Cleveland Jr., J. C., Fazzalari, F. L., Filardo, G., Normand, S-L. T., Furnary, A. P., Magee, M. J., Rankin, J. S., Welke, K. F., Han, J., & O'Brien, S. M. (2015). The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: A report of the Society of Thoracic Surgeons Quality Measurement Task Force. *The Annals of Thoracic Surgery, 10*(4), 1315-1325. https://doi.org/10.1016/j.athoracsur.2015.06.122

Stulberg, J.J., Delaney, C.P., Neuhauser, D.V., Aron, D.C., Fu, P., & Koroukian, S.M. (2010). Adherence to
    surgical care improvement project measures and the association with postoperative infections.
    JAMA, 303(24), 2479–2485. https://doi.org/10.1001/jama.2010.841