Committee Guide to NQF's Measure Endorsement Process

LAST UPDATED: JANUARY 9, 2014

Contents

I. The National Quality Forum

Who is NQF?

The National Quality Forum (NQF), established in 1999, is a non-profit, non-partisan, membership-based organization that is recognized and funded in part by Congress and entrusted with the important public service responsibility of bringing together various public and private sector organizations to reach consensus on how to measure quality in healthcare as the nation work to make it better, safer, and more affordable.

NQF was created by a coalition of public- and private-sector leaders in response to the recommendation of the *Advisory Commission on Consumer Protection and Quality in the Health Care Industry*. In its final report, published in 1998, the commission concluded that an organization like NQF was needed to promote and ensure patient protections and healthcare quality through measurement and public reporting.

Who is involved at NQF?

NQF has 440 organizational members who give generously of their time and expertise. In 2012, more than 822 individuals volunteered on more than 41 NQF-convened committees, working groups, and partnerships. We estimate that this time conservatively translates into more than 55,000 hours or $4 million donated to NQF efforts in 2012, which reflects true commitment to the quality cause. The NQF Board of Directors governs the organization and is composed of key public- and private-sector leaders who represent major stakeholders in America's healthcare system. Consumers and those who purchase healthcare hold a simple majority of the at-large seats.

Member organizations of NQF have the opportunity to take part in a national dialogue about how to measure healthcare quality and publicly report the findings. Members participate in NQF through one of eight Member Councils:

- Consumer Council
- Health Plan Council
- Health Professionals Council
- Provider Organizations Council
- Public/Community Health Agency Council
- Purchasers Council
- Quality Measurement, Research, and Improvement Council
- Supplier and Industry Council

Each of these councils provides unique experiences and views on healthcare quality that are vital to building broad consensus on improving the quality of healthcare in America. Together, NQF members promote a common approach to measuring and reporting healthcare quality and fostering system-wide improvements in patient safety and healthcare quality. NQF's membership spans all those interested in healthcare. Consumers and others who purchase healthcare sit side-by-side with those who provide care and others in the healthcare industry. Expert volunteers and members are the backbone of NQF work.

What does NQF do?

Ten years ago, working with all major healthcare stakeholders, NQF endorsed its first voluntary, national consensus performance measures to answer the call for standardized measurement of healthcare services. After 10 years, we have a portfolio of more than 600 NQF-endorsed™ measures –most of which are in use by both private and public sectors, and an enormous body of knowledge about measure development, use, and performance improvement. NQF plays a key role in shaping our national health and healthcare improvement priorities, including the National Quality Strategy, through its convening of the National Priorities Partnership. NQF also provides public input to the federal government and the private sector on optimal, aligned measure use via its convening of the Measures Application Partnership.

NQF reviews, endorses, and recommends use of standardized healthcare performance measures. Performance measures are essential tools used to evaluate how well healthcare services are being delivered. NQF's endorsed measures often are invisible at the clinical bedside but quietly influence the care delivered to millions of patients every day. Performance measures can:

- make our healthcare system more information rich;
- point to actions physicians, other clinicians, and organizations can take to make healthcare safe and equitable;
- enhance transparency around quality and cost of around quality and cost of healthcare;
- ensure accountability of healthcare providers; and
- generate data that helps consumers make informed choices about their care.

Working with members and the public, NQF also helps define our national healthcare improvement 'to- do' list, and encourages action and collaboration to accomplish performance improvement goals.

Who benefits from this work?

Standardized healthcare performance measures help clinicians and other health care providers understand whether the care they provided their patients was optimal and appropriate, and if not, where to focus their efforts to improve the care they deliver. Measures are also used by all types of public and private payers for a variety of accountability purposes, including public reporting and pay-for- performance. Measures are an essential part of making quality and cost of healthcare more transparent to all, importantly for those who receive care or help make care decisions for loved ones. Use of standardized healthcare performance measures allows for comparison across clinicians, hospitals, health plans, and other providers.

Where do I find NQF-endorsed™ measures?

The Quality Positioning System (QPS) is a web-based tool that helps you more easily find NQF-endorsed® measures. Search by measure title or number, as well as by condition, care setting, or measure steward. Driven by feedback from users, QPS 2.0 now allows users to search for measures by their inclusion in Federal reporting and payment programs; to provide feedback any time about the use and usefulness of measures; and to view measures that are no longer NQF-endorsed. QPS can also be used to learn from other measure users about how they select and implement measures in their performance improvement programs. The QPS may be accessed at this link.

Where do I find more information about NQF?

The Field Guide to NQF Resources is a dynamic, online resource designed to help those involved with measurement and public reporting more easily access basic information and NQF resources related to performance measurement.

Glossary of Terms

A comprehensive glossary of terms used in NQF activities as well as performance measurement and quality improvement in general can be found on the NQF web site on the Submitting Standards page.


NATIONAL QUALITY FORUM
AN ILLUSTRATIVE EXAMPLE
Lifecycle of a Performance Measure:
Depression Remission at 6 months
1
PREVALENCE OF DISEASE
The (APA) American Psychiatric Association has data that show 1 in 10 are depressed. There are evidence-based treatments that can lead to remission of symptoms.
2
ASSESSMENT TOOL
An available standardized tool is used to assess prevalence and severity of depression for a given population.
3
LOCAL INITIATIVE
MN Community Measurement developed and tested a way to measure whether a patient's depression is in remission 6 months after treatment.
MINNESOTA
MEASURE
1
2
3
NATIONAL CONSENSUS STANDARD
6
NQF endorsed the measure as a national consensus standard.
USA
MN
RESULTS MADE PUBLIC
5
MN Health Scores website publicly reports local performance on depression remission.
MINNESOTA HealthScores
5%
7%
33%
18%
12%
25%
RESULTS SPUR CHANGE IN PRACTICES
4
The Institute for Clinical Systems improvement helped doctors implement change in their practices that lead to improved results.
Doctor's Office
6
5
4
7
ELECTRONIC HEALTH RECORDS
MN Community Measurement retooled the measure for use in electronic health records.
Jane Smith
Age: 47 0085974
Date of Last
PHQ-9: 12/1/11
MINNESOTA HealthScores
8
HIT PAYMENT QUALIFICATION
Depression improvement at 6 months was suggested for inclusion in CMS' Meaningful Use HIT payment program by an NQF convened group, eventually leading to more widespread adoption and improvement in patient care.
7
8
IMPROVE
Source: National Quality Forum, MN Community Measurement 10-2012
Illustration: Funnel, Inc.
WWW.QUALITYFORUM.ORG

II. The Evolving Performance Measurement Landscape

For more than a decade the quality measurement enterprise – the many organizations focused on performance measurement to drive improvement in the quality and cost of healthcare provided in the United States – has rapidly grown to meet the needs of a diverse and demanding market place. As a result of greater experience with measurement stakeholders have identified priorities for certain types of performance measures:

Outcome measures —Stakeholders are increasingly looking to outcome measures because the end results of care are what matter to everyone. Outcome measures assess rates of mortality, complications, improvement in symptoms or functions. Outcome measures, including patient experiences and patient-reported outcomes, seek to determine whether the desired results were achieved. Measuring performance on outcomes encourages a "systems approach" to providing and improving care.

Composite measures —Composite performance measures, which combine information on multiple individual performance measures into one single measure, are of increasing interest in healthcare performance measurement and public accountability applications. According to the Institute of Medicine, such measures can enhance the performance measurement enterprise and provide a potentially deeper view of the reliability of the care system.

Measures over an episode of care—To begin to define longitudinal performance metrics of patient-level outcomes, resource use, and key processes of care NQF has endorsed a measurement framework for patient-focused episodes of care. This framework proposes a patient-centered approach to measurement that focuses on patient-level outcomes over time—soliciting feedback on patient and family experiences; assessing functional status and quality of life; ensuring treatment options are aligned with informed patient preferences; and using resources wisely.

Measures that address healthcare disparities—NQF has established a broader platform for addressing healthcare disparities and cultural competency by identifying a set of disparities-sensitive measures among the existing NQF portfolio of endorsed measures. These disparities-sensitive measures should be routinely stratified and reported by race/ethnicity and language. Additionally, the disparities-sensitive criteria were finalized and incorporated into a prospective approach for the assessment of disparities-sensitivity for all new and maintenance measures submitted to NQF.

Measures that are harmonized —The current quality landscape contains a proliferation of measures, including some that could be considered duplicative or overlapping, and others that measure similar but measure the same concepts and/or patient populations somewhat differently. Such duplicative measures and/or those with similar but not identical specifications may increase data collection burden and create confusion or inaccuracy in interpreting performance results for those who implement and use performance measures. Recognizing that NQF can take on more of a facilitator role while accounting for the needs of measure developers, NQF has proposed a revised process to ensure harmonization and competing measures issues are adequately addressed and provide adequate time to develop to resolve questions.

Measures for patients with multiple chronic conditions—Under the direction of the multi-stakeholder Multiple Chronic Conditions (MCCs) Standing Committee, NQF has developed a person-centric measurement framework for individuals with MCCs. Specifically, this framework provides a definition for MCCs, identifies high- leverage domains for performance measurement, and offers guiding principles as a foundation for supporting the quality of care provided to individuals with MCCs.

eMeasures and Health Information Technology (HIT)—NQF is committed to improving healthcare quality through the use of health information technology (IT). Care can be safer, more affordable, and better coordinated when electronic health records (EHRs) and other clinical IT systems capture data needed to measure performance, and when that data are easily shared between IT systems. Our health IT initiatives – made up of several distinct yet related areas of focus – are designed to support an electronic environment based on these ideals; more importantly, they are designed to help clinicians improve patient care.

The National Quality Strategy (NQS)

The Department of Health and Human Services' (HHS) release of the first National Quality Strategy (NQS) in 2011 marked a significant step forward in the effort to align a healthcare system characterized by intense fragmentation. The NQS' aims and goals set forth a unified vision of the healthcare system that was understandable and applicable to all stakeholders at every level—local, state, and national.

The National Quality Strategy—heavily informed by the NQF-convened, private-public National Priorities Partnership—laid out a series of six priorities for focusing the nation on how to best and most rapidly improve our health and healthcare. NQF has carefully aligned its work with these goals, utilizing them as a roadmap for much of its work. Currently, NQF-endorsed measures are being tagged to the NQS.

The "triple aims" of the National Quality Strategy will be used to guide and assess local, State, and national efforts to improve health and the quality of health care:

- Better Care: Improve the overall quality, by making health care more patient-centered, reliable, accessible, and safe.
- Healthy People/Healthy Communities: Improve the health of the U.S. population by supporting proven interventions to address behavioral, social and, environmental determinants of health in addition to delivering higher-quality care.
- Affordable Care: Reduce the cost of quality health care for individuals, families, employers, and government.

To advance these aims, the National Quality Strategy will focus initially on six priorities:

- Making care safer by reducing harm caused in the delivery of care.
- Ensuring that each person and family is engaged as partners in their care.
- Promoting effective communication and coordination of care.
- Promoting the most effective prevention and treatment practices for the leading causes of mortality, starting with cardiovascular disease.

- Working with communities to promote wide use of best practices to enable healthy living.
- Making quality care more affordable for individuals, families, employers, and governments by developing and spreading new health care delivery models.

III. The ABCs of Measurement

According to the Institute of Medicine (IOM) definition, a performance measure is the "numeric quantification of healthcare quality." IOM defines quality as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge." Thus, performance measures can quantify healthcare processes, outcomes, patient perceptions, and organizational structure and/or systems that are associated with the provision of high-quality care.

Performance measures are widely used throughout the healthcare arena for a variety of purposes. Not all measures are suitable for NQF's dual purpose of accountability (including public reporting) and performance improvement. NQF does not endorse measures intended only for internal quality improvement.

NQF's **ABCs of Measurement** brochure describes various aspects of performance measurement:

- The Difference a Good Measure Can Make
- Choosing What to Measure
- The Right Tools for the Job
- Patient-Centered Measures = Patient-Centered Results
- What NQF Endorsement Means
- How Endorsement Happens
- How Measures Can Work: Safety
- How Measures Will Serve Our Future
- What You Can Do

NATIONAL QUALITY FORUM

Understanding Performance Measures: Anatomy and Types

WHAT IS A PERFORMANCE MEASURE?

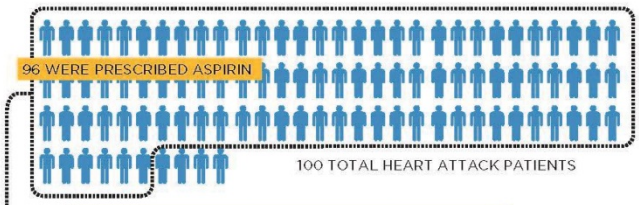
A healthcare performance measure is a way to calculate whether and how often the health and healthcare system does what it should.

Measures are based on scientific evidence about **processes, outcomes, perceptions, or systems** that relate to high-quality care.

Health and Healthcare System

CONSTRUCTING A MEASURE

The result of a measure is usually shown as a ratio or a percentage, and allows for comparison to other providers and benchmarking against national and local performance.



MEASURE FORMULA

NUMERATOR: # WHO HAD A SPECIFIC TREATMENT / # ELIGIBLE FOR TREATMENT (DENOMINATOR) = % (RESULT)

MEASURE EXAMPLE

96 HEART ATTACK PATIENTS WERE APPROPRIATELY PRESCRIBED ASPIRIN AT DISCHARGE / 100 TOTAL HEART ATTACK PATIENTS = 96%

EXAMPLE: Once a person has had a heart attack, taking aspirin daily has been shown to reduce the chance of having a second one. Guidelines tell physicians to prescribe aspirin to all patients leaving the hospital after treatment.

TYPES OF PERFORMANCE MEASURES

STRUCTURAL MEASURES

ASSESS HEALTHCARE INFRASTRUCTURE

EXAMPLE: The percentage of physicians in a practice who have systems to track and follow patients with diabetes.

PROCESS MEASURES

ASSESS STEPS THAT SHOULD BE FOLLOWED TO PROVIDE GOOD CARE

EXAMPLE: The percentage of patients with diabetes who have had an annual eye exam in the last year.

OUTCOME MEASURES

ASSESS THE RESULTS OF HEALTHCARE THAT ARE EXPERIENCED BY PATIENTS

EXAMPLE: The percentage of diabetes patients who are blind or have compromised vision.

08-2013
Illustration: Funnel, Inc.

WWW.QUALITYFORUM.ORG

IV. NQF Endorsement of Consensus Standards

How does NQF endorse measures?

NQF uses a formal Consensus Development Process (CDP) to evaluate and endorse consensus standards, including performance measures, best practices, frameworks, and reporting guidelines. The CDP is designed to call for input and carefully consider the interests of stakeholder groups from across the healthcare industry.

Because NQF uses this formal process, it is recognized as a voluntary consensus standards-setting organization as defined by the National Technology Transfer and Advancement Act of 1995 and Office of Management and Budget Circular A-119.

Over the past 10 years, the processes that form NQF's CDP and its implementation have evolved to ensure that evaluation of candidate consensus standards continues to follow best practices in performance measurement and standards-setting.

NQF's Consensus Development Process involves eight principal steps. Each contains several sub- steps and is associated with specific actions. The steps are:

1. Call for Nominations - Transitioning to Standing Committees

NQF strives to continually improve its measure endorsement process so as to remain responsive to its stakeholders' needs. Volunteer, multi-stakeholder committees are the central component to this process, and the success of NQF's projects is due in large part to the participation of its Steering Committee members.

Composition of Standing Committees

Standing topical Committees will include 20 individuals with the option to flex up to 25 individuals if include specialized expertise is needed, after consultation with the NQF membership. The Standing Committee will represent a variety of stakeholders, including consumers, purchasers, providers, health professionals, health plans, suppliers and industry, community and public health, and healthcare quality experts. Because NQF attempts to represent a diversity of stakeholder perspectives on committees, a limited number of individuals from each of these stakeholder groups can be seated on a committee.

Nominations are to an individual, not an organization, so "substitutions" of other individuals from an organization during conference calls or meetings are not permitted. Committee members are encouraged to engage colleagues and solicit input from colleagues throughout the process.

Standing Committee Terms

During the transition from project-specific Steering Committees to Standing Committees, committee members will be appointed to a two or three year term initially, with approximately half of the committee appointed to a two year term and the other half a three year term. Each term thereafter will be a three year term. Committee members may serve two consecutive terms. They must step down for a full term (three years) before becoming eligible for reappointment. The Committee member's term on the Standing Committee begins upon selection to the Committee, immediately following the close of the roster commenting period.

Evolution of Standing Committees

- *Prior to the HHS contract that started in 2009, NQF operated with a great deal of uncertainty regarding resources for proposed projects. Consequently, measure endorsement work was organized on a project-by-project basis with no comprehensive schedule. NQF appointed project-specific Steering Committees, with the nominations process commencing when project funding had been secured.*
- *NQF established a three-year schedule for Endorsement Maintenance projects across 20 cross-cutting and condition-specific areas. NQF is currently in the process of convening a set of Standing Committees within various project topic areas. Committee members will initially serve two or three year terms, and the Committees will be responsible for handling endorsement of both new and maintenance measures, as well as ad hoc and expedited project work in their designated areas.*

Standing Committee expectations and time commitment

Participation on the Committee requires a significant time commitment. To apply, Committee members should be available to participate in all currently scheduled calls/meetings. Over the course of the Committee member's term, additional calls will be scheduled or calls may be rescheduled; new dates are set based on the availability of the majority of the Committee.

Committee participation includes: (these times may vary depending on the number and complexity of the measures under review as well as the complexity of the topic and multi-stakeholder consensus process)

- Review all measure submission forms (approximately 1-2 hours per measure)
- Participate in the scheduled orientation call (2 hours)
- Complete all surveys and evaluations
- Review measures on workgroup calls (2 hours); workgroup assignments will be made by area of expertise
- Attend scheduled in-person meetings (2 full days in Washington, DC); in-person meetings typically will take place on an annual basis
- Complete measure review by participating on the post-comment conference call (2 hours)
- Complete additional measure reviews by conference call if needed;

- Participate in additional calls as necessary
- Present measures and lead discussions for the Committee on conference calls and in meetings

If a member has poor attendance or participation:

- The NQF staff will contact the member asking if he/she would like to forego their Committee participation.

If a member is unable to fulfill their term (for any reason):

- The nominations received during the most recent call for nominations would be reviewed for a replacement.
- NQF staff will contact the potential replacement.
- Upon acceptance of committee appointment, the new committee member would complete the term of the individual they have replaced.
- The out-going member may not select a substitute to carry out the remainder of the term.

Standing Committee Disclosure of Interest

Per the NQF Disclosure of Interest Policy for CDP Standing Committees, each nominee will be asked to complete a general disclosure of interest (DOI) form for each Committee to which they have applied prior being seated on the Committee. The DOI form for each nominee is reviewed in the context of the topic area in which the Committee will be reviewing measures.

Once nominees have been selected to serve on the Committee , during the 14-day roster comment period a measure-specific DOI form will be distributed to determine whether any , member(s) will be required to recuse themselves from discussion of one or more measures under review based on prior involvement or relationships to entities relevant to the topic area.

As a member of an NQF Standing Committee, you will be asked to review various types of measures throughout your term; you may be asked to complete this form for each batch of measures under review by the Committee to ensure any potential conflicts or biases have been identified.

Standing Committee application requirements

Self-nominations are welcome. Third-party nominations must indicate that the individual has been contacted and is willing to serve. To be considered for appointment to the Standing Committee, please send the following information:

- a completed online nomination form, including:
    - a brief statement of interest
    - a brief description of nominee expertise highlighting experience relevant to the committee
    - a short biography (maximum 100 words), highlighting experience/knowledge relevant to the expertise described above and involvement in candidate measure development

- curriculum vitae or list of relevant experience (e.g., publications) *up to 20 pages*

- a completed electronic disclosure of interest form. This will be requested upon your submission of the nominations form for Committees actively seeking nominees.

- confirmation of availability to participate in currently scheduled calls and meeting dates.

Materials should be submitted through the NQF website.

2. Call for Candidate Standards (Measures or Practices)

Before the start of a project, NQF issues a formal call for candidate standards. Each candidate measure will have a measure steward who assumes responsibility for the submission of the measure for potential endorsement to NQF. Both new and maintenance measures are accepted in this call. The measure steward is responsible for making the necessary updates to the measure, informing NQF about any changes that are made to the measure on an annual basis, and providing the required measure information for the measure maintenance process that occurs approximately every three years. To submit a measure for an initial endorsement evaluation or a maintenance review, a measure steward must complete and submit specific information about the measure via an online form through the NQF website.

3. Candidate Consensus Standards Review

The relevant committee conducts a detailed review of all submitted standards, sometimes with the help of a technical advisory panel. The duration of a Committee's review of the candidate consensus standards for a given project can vary depending on the scope of the project, the number of standards under review, and the relative complexity of the standards.

During this review process, the committee may meet several times, via conference calls and/or in- person meetings, to discuss and evaluate the submitted consensus standards in accordance with NQF criteria and guidance. All meetings and conference calls of a committee and any associated technical advisory panel(s) are open to NQF members and the public. Information about each of these meetings, including the agenda and the location or dial-in information, is posted on NQF's public website, through both the events calendar and the specific webpage for the project. Each meeting or conference call of a Standing committee includes a specific period during which NQF members and interested members of the public may make comments regarding the committee's deliberations.

Details of the review are described in Section V of this guidebook.

4. Public and Member Comment

After a committee completes its initial review of the submitted candidate standards, a draft of the committee's recommendations--or "draft report"-- is posted on the NQF website for review and comment by members of NQF and the public. Both NQF members and interested members of the public can submit comments on the standing committee's draft recommendations through the NQF website. As part of NQF's commitment to transparency, all submitted comments will be posted on the NQF website, where they can be reviewed by any site visitor.

All submitted comments are reviewed by the standing committee, and all submitted comments receive responses from the Standing committee, measure developers, and/or NQF, as appropriate. The Standing committee may revise its recommendations in direct response to a specific comment or series of comments that are submitted during this phase of the CDP.

5. NQF Member voting

Once a committee has reviewed all of the comments submitted during the public and member commenting period and made any desired revisions to their recommendations, members of NQF vote on the candidate standards that are recommended by the committee.

6. Consensus Standards Approval Committee (CSAC) Decision

The work of the CSAC focuses on the approval of proposed consensus standards and the ongoing enhancement of NQF's CDP. Members of the CSAC possess breadth and depth of expertise and are drawn from a diverse set of healthcare stakeholders with a simple majority of consumers and purchasers. Some CSAC members possess specific expertise in measure development, application, and reporting.

The CSAC reviews the recommendations of standing committees, the comments received, and the results of the NQF Member vote. After detailed review of a candidate standard, the CSAC determines if consensus has been reached across the various NQF Member Councils. They seek further input from Council Leaders if there is a lack of consensus. On some occasions, the CSAC may also request a second round of Member voting on a particular candidate standard or set of standards. The CSAC can recommend full endorsement, time-limited endorsement, or recommend against endorsement of a candidate standard.

The CSAC also serves in an advisory capacity to the Board of Directors and NQF management on ongoing enhancements to the Consensus Development Process and emerging issues in performance measurement.

7. Board Ratification

CSAC decisions regarding consensus standards are submitted to NQF's Board of Directors. The Board can affirm or deny a CSAC decision. All consensus standards that are approved by the CSAC must be ratified by the Board for endorsement.

8. Appeals

After a consensus standard has been formally endorsed by NQF, any interested party may file an appeal of the endorsement decision with the NQF Board of Directors within 30 days. An appeal may only be filed in response to NQF *endorsement* of a candidate standard or set of standards; that is, an interested party may not file an appeal regarding the decision to not endorse a candidate standard.

NATIONAL QUALITY FORUM

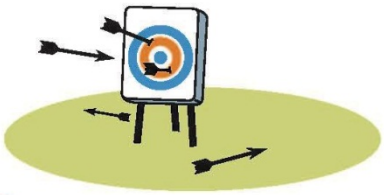
Multi-stakeholder Review: Criteria for Evaluating a Measure

MULTI-STAKEHOLDER COMMITTEES OVERSEE ENDORSEMENT

These committees evaluate measures by clinical condition against agreed upon criteria. Measures reviewed are endorsed and receive the NQF seal of approval. In order to receive NQF endorsement, measures must meet all five endorsement criteria.

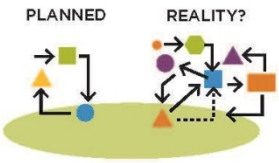
1 IMPORTANCE TO MEASURE AND REPORT

Evaluate whether the measure has potential to drive improvements in care, is aligned with the National Quality Strategy, and is based on strong clinical evidence.



2 SCIENTIFIC ACCEPTABILITY OF MEASURE PROPERTIES

Determine if the measure will allow for valid conclusions about quality based on performance scores. If measures are not reliable (consistent) and valid (correct), results may mis-classify providers.



3 FEASIBILITY

Assess the burden involved with collecting measure information.

4 USABILITY AND USE

Evaluate if the measure can be appropriately used in accountability and improvement efforts.

5 ASSESS RELATED AND COMPETING MEASURES

Determine whether the measure is duplicative of other measures. If other criteria are met, harmonize or select the best measure among duplicative measures.

ACTION

08-2013
Illustration: Funnel, Inc.

WWW.QUALITYFORUM.ORG

V. The Measure Evaluation Process

Role of the Standing Committee

The Standing Committee acts as a proxy for the NQF multi-stakeholder membership. The individual members of the Committee are selected from the various stakeholder groups. *Each committee member is expected to participate as an individual and not as a representative of any specific organization.* Although individuals may wear "many hats" with different points of view, committee members should use their own personal experience and expertise while serving on the committee.

The primary responsibility of the Committee is to evaluate the candidate measures using NQF's standard measure evaluation criteria. The Committee also will consider the National Quality Strategy and NQF's frameworks to review the entire portfolio when making recommendations for endorsement of individual measures in a topic area and identify measure gaps.

A document containing a short biography of all Standing Committee members is posted on the NQF project webpage and on the project SharePoint site. Committee members are encouraged to review the bios as you get to know your fellow Committee members.

Standing Committee expectations

- *Attend all meetings and conference calls;*
- *Identify and disclose potential biases (real or perceived);*
- *Review assigned measures using NQF evaluation criteria and guidance;*
- *Lead discussion of some measures at calls or meetings;*
- *Participate in the discussion and vote on ratings and recommendations for all measures;*
- *Review meeting summaries and draft reports;*
- *Review public comments and suggest responses.*

Role of the Committee co-chairs

Typically, two Committee members are selected to serve as co-chairs. The co-chairs' responsibilities are to:

- Facilitate Standing Committee calls and meetings;
- Work with NQF staff to achieve the goals of the project;
- Assist NQF staff in anticipating questions and identifying additional information that may be useful to the Committee;
- Participate as a full voting member of the Standing Committee;
- Represent the Committee at the CSAC meetings or calls.

Standing Committee Process for Evaluating and Recommending Measures

Measures submitted for review

Measure stewards/developers submit measures for consideration in a standardized form that is structured to solicit the information necessary for committees to determine whether the NQF criteria are met. The submission form—which is comprised of an online form and two MS Word attachments— is posted on NQF's web site for transparency.

Measure evaluation criteria

NQF endorses performance measures that are suitable for both accountability applications (e.g., public reporting, accreditation, performance-based payment, network inclusion/exclusion, etc.) and internal quality improvement efforts. NQF's measure evaluation criteria and subcriteria are used to determine the suitability of measures for use in these activities. Because endorsement initiates processes and infrastructure to collect data, compute performance results, report performance results, and improve and sustain performance, NQF endorsement is intended to identify those performance measures that are most likely to facilitate achievement of high quality and efficient healthcare for patients.

SharePoint site

- *Standing Committee members will receive the access link and password for the project SharePoint site.*
- *All project documents will be housed on SharePoint to provide ready access for all members.*
- *SharePoint also has a discussion platform that can be used to conduct offline discussions of project or measure issues.*
- *SharePoint has a survey tool that will be used to collect information on the initial reviews.*
- *If you have difficulty accessing the SharePoint site please contact the NQF project staff.*

To determine whether a candidate measure should be endorsed by NQF, the standing committee evaluates the candidate measures against NQF's standard measure evaluation criteria. These criteria have evolved over time to reflect the input of a wide variety of stakeholders and the needs indicated by those stakeholders for the measures that will hold people accountable for the care that they deliver. The standard criteria foster consistency and predictability for measure developers and for those using NQF-endorsed measures.

Committee members are expected to familiarize themselves with the criteria and use the criteria to make recommendations for endorsement. NQF staff will provide an initial review of the measure information to assist the committee in its evaluation.

NQF's criteria are organized around five major concepts with subcriteria that further describe how the main criteria are demonstrated. The criteria are arranged in a hierarchy for review and evaluation.

The main criteria and rationale for order of evaluation follow below. More detail regarding the subcriteria and evaluation are provided in subsequent sections of this guidebook.

Main criteria

- **Importance to Measure and Report** (this is not the same as "Important to do") - Extent to which the specific measure focus is evidence-based, important to making significant gains in healthcare quality, and improving health outcomes for a specific high-priority (high-impact) aspect of healthcare where there is variation in or overall less-than-optimal performance.

  This is a must-pass criterion. If a measure does not meet the importance criterion, then the other criteria are less meaningful.

- **Reliability and Validity: Scientific Acceptability of the Measure Properties** - Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of care when implemented.

  This is a must-pass criterion. The goal of measuring performance is to make valid conclusions about quality; if a performance measure is not reliable and valid, there is a risk of misclassification and improper interpretation.

- **Feasibility** - Extent to which the specifications, including measure logic, require data that are readily available or could be captured without undue burden and can be implemented for performance measurement.

  Ideally, performance measurement should create as little burden as possible; however, if an important and scientifically acceptable measure is not feasible, alternative approaches and strategies to minimize burden should be considered.

- **Usability and Use** - Extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations.

  NQF-endorsed measures are intended to be used for decisions related to accountability and improvement. New measures should have a credible plan for implementation in accountability applications and rationale for use in improvement. Measures undergoing endorsement maintenance are expected to be in use.

- **Comparison to Related and Competing Measures** - If a measure meets the above criteria and there are endorsed or new related measures (either the same measure focus or the same target population) or competing measures (both the same measure focus and the same target population), the measures are compared to address harmonization and/or selection of the best measure.

  Duplication and lack of harmonization among performance measures create burdens related to inefficient use of resources measure development, increased data reporting requirements, and confusion when they produce conflicting results.

For each of the standard criteria, several sub-criteria delineate how to demonstrate that the major criteria are met (i.e., how do you know a measure is important, scientifically acceptable, etc.?). NQF's criteria

parallel best practices for measure development (for example: begin with identifying what is important to measure, and later what is feasible). Most criteria/subcriteria involve a matter of degree rather than all-or-nothing determination – this requires both evidence and expert judgment. The measure evaluation criteria will be discussed in more detail in Section VI.

Committee members first review and evaluate the measures individually and in workgroups, but ultimately the entire Standing Committee as a whole determines—for each measure—to what extent the criteria are met and whether to recommend the measure for NQF endorsement. NQF recognizes that each committee member brings different expertise and experience to the project and may not feel qualified to evaluate all aspects of a measure. All committee members should contribute to the evaluation to the best of his/her ability, knowing that the final evaluation rating and recommendation will be made by the full Standing Committee.

Initial Measure Evaluation

- *Each workgroup will be assigned 4- 6 measures for review.*
- *Plan to spend at least 1 hour per measure though you may need less time as you gain familiarity and experience with NQF's measure evaluation criteria.*
- *Lead discussants will likely spend more time on their assigned measures.*

NQF staff review and evaluation work sheet

To assist Committee, an evaluation worksheet will be placed on top of the Measure Information Form submitted by the measure developer/steward. NQF staff will review the submitted information against the NQF criteria and highlight areas for specific discussion by the committee as well as any specific questions or critical decisions for the committee to consider. The worksheet will include any feedback from the field on implementation or use of the measure as well as measure specific comments submitted to NQF. Internal links will be used to navigate through the document. Use CTRL + click to navigate to a link and ALT + LEFT (left arrow) to return.

Initial evaluation by individual committee members and workgroups

In order to ensure an in-depth evaluation of all measures, the Standing Committee may be divided into workgroups that will focus on a subset of the measures being considered in the project. Workgroups will meet by conference call for preliminary discussion of the measures and how well they meet the evaluation criteria. Measure developers generally attend the calls and are available to answer questions or make clarifications regarding their measures. The workgroup calls are open to the public. Committee members who are assigned to a particular workgroup are expected to review all measures assigned to that workgroup in detail and to participate in the workgroup call. All committee members are encouraged to review all measures in detail and are welcome to attend all workgroup calls.

When conducting the initial in-depth evaluations, each Committee member will consider all assigned measures in light of all criteria and subcriteria prior to the workgroup calls. A SharePoint survey tool will be provided to collect your initial thoughts about the measures for further discussion. ***If you have difficulty using the SharePoint survey tool, please let us know so that we can assist you.***

Lead discussants

To facilitate the Committee discussions, 1-2 lead discussants will be designated for each measure. These lead discussants will:

- be fully conversant with the submitted measure information on the assigned measures;
- evaluate the assigned measures against the NQF measure evaluation criteria and submit comments prior to the workgroup call;
- begin the discussion of the measure evaluation including:
  - presenting a brief description of the measure
  - summarizing the evaluation of each criterion based on all the workgroup's pre-call evaluation comments (these comments be made available to you prior to the workgroup calls), highlighting areas of concern or difference of opinion and the issues or questions posed in the staff review;
  - verbalizing conclusions regarding how well the measure meets NQF's evaluation criteria (refer to Section VI on the criteria).

The discussion points raised during the workgroup call will be added to the staff summary document in preparation for the in-person meeting.

Due to the large volume of documents for the meeting, Committee members are requested to bring laptop computers and view the documents electronically.

Internet connection is available.

Evaluation by the entire committee at the in-person meeting or web meeting

NQF Standing Committee will meet either in-person or by web meetings to evaluate measures and make recommendations.

- **Transparency** - The meeting/web meeting is open to the public (in-person and by phone). The proceedings are transcribed and posted on NQF's web site.
- **Disclosure of Interests** - During introductions at the beginning of the meeting, each Committee member is asked to disclose any interests as identified on your Disclosure of Interest form.
- Measure developers will be present during the meeting (in person or via phone) to respond to any issues or questions. Measure developers are given an opportunity to provide a brief introduction to their measures at the beginning of each topic area. The discussion surrounding the evaluation of the measures is meant primarily for the committee members. However, the committee may consult measure developers to clarify information about the measure or explain various decisions regarding measure development.
- Each measure is evaluated individually by the Standing Committee. The lead discussant will introduce each measure and begin the discussion. After discussion by the entire Committee, a vote is taken on the each criterion and selected subcriteria, and finally, on whether the measure

meets the NQF criteria to be recommended for endorsement. The entire Standing Committee determines to what extent the criteria are met for each measure and whether to recommend measures for endorsement. Related and competing measures are addressed only if measures are considered suitable for endorsement.

- During measure evaluation Committee members often offer suggestions for improvement to the measures. These suggestions can be considered by the developer for future improvements; however, the Committee is expected to evaluate and make recommendations on the measures per the submitted specifications and testing.

- Voting by the Standing Committee – A measure is recommended for endorsement by the Standing Committee when the vote margin on all major criteria (Importance, Scientific Acceptability) and overall is greater than 60% of voting members in favor of endorsement. A measure is not recommended for endorsement when the vote margin on any major criteria or overall is less than 40% of voting members in favor of endorsement. The Standing Committee has not reached consensus if the vote margin on any major criterion or overall is between 40%-60% in favor of endorsement.

    - When the Standing Committee has not reached consensus, all measures for which consensus was not reached will be put out for NQF Member and public comment. The Standing Committee will consider the comments and re-vote on measures where consensus was not reached. After the re-vote, all measures that are recommended (>60% in favor of endorsement) by the Standing Committee or where consensus has not been reached (between 40%-60% in favor of endorsement) will be put out for NQF Member vote.

- NQF Members and the public are provided opportunities to comment at designated times during the meeting.

Committee ground rules for workgroup calls and meetings

Committee members act as a proxy for NQF's membership. As such, this multi-stakeholder group brings varied perspectives, values, and priorities to the discussion. Respect for differences of opinion and collegial interactions with other committee members and measure developers are critical.

The workgroup call and in-person meeting agendas are typically quite full. All Committee members are responsible for ensuring that the work of the meeting is completed during the time allotted. During these discussions, Committee members should:

- fully disclose all potential biases or interests in the measures under discussion;
- be prepared, having evaluated the measures beforehand;
- base evaluation and recommendations on the measure evaluation criteria and guidance;
- remain engaged in the discussion without distractions;
- not leave the meeting/call except at breaks;

- keep comments concise and focused;
- avoid dominating a discussion and allow others to contribute; and
- indicate agreement without repeating what has already been said.

After the in-person meeting

After a project's Standing Committee completes its initial review of the submitted measures, a draft of the Committee's recommendations--or "draft technical report"-- is posted on the NQF website for review and comment by members of NQF and the public. All measures evaluated in the project, regardless of the recommendation, are posted for public and member comment.

NQF Member and Public comment period

When a comment period opens, a notification is posted on the NQF website, and will be available through the event calendar and on the specific project page. NQF also sends out an email notification to NQF members and members of the public who have signed up for these notifications. Both NQF members and interested members of the public can submit comments on the Standing Committee's draft report via the NQF website. As part of NQF's commitment to transparency, all submitted comments will be posted on the NQF website, where they can be reviewed by any site visitor.

The 30-day comment period serves to enable feedback to the Standing Committee on their evaluation and recommendations for endorsement. NQF Members and non-members value the opportunity to weigh in on the deliberations, often offering constructive criticism, alternative viewpoints, or support for the Committee's recommendations. The comments are available for viewing during the comment period. Committee members are welcome to check the comments throughout the comment period. An important responsibility for the Committee is responding to the comments***. The Committee is expected to thoughtfully consider the comments and adjust any recommendations as needed.***

<u>Developer request for reconsideration of a measure not recommended</u>

Requests for reconsideration related to appropriate application of the criteria are submitted through the public and member comment process. The request must cite the specific evaluation criteria or subcriteria that the developer thinks was not applied properly to the specific information as submitted and evaluated by the Standing committee. The Standing committee will review the cited information in the submission form and criteria under question during the comment review process, with the option to re-vote on the measure.

Post-comment conference call

After the conclusion of the member and public comment periods, the Standing Committee meets by conference call to review all submitted comments. The Standing Committee may also seek out technical advice or other specific input from external sources, as needed. Measure developers may be invited to respond to comments, particularly if the comment relates to the specifications of the measure.

After its review of the submitted comments, the Standing Committee may choose to revise its initial recommendations in response to a specific comment or series of comments. Any revisions will be reflected in a revision of the draft report.

Should the Standing Committee determine its revisions to be substantial in nature, the revised version of the draft report may be re-circulated for a second comment period for members and the public. If a revised version of the draft report is re-circulated for a second comment period, the review will follow the same process as the initial review and comment period.

NQF member voting

Once a Standing Committee has reviewed all of the comments submitted during the public and member comment period and made any revisions to the draft report, NQF members may vote on the measures that are recommended by the Committee (>60% of the Committee members vote in favor of endorsing the measure) or for which the Committee has not reached consensus (between 40%-60% of the Committee members vote in favor of endorsing the measure).

All NQF member organizations are eligible to vote on any consensus development project. Each voting period is open for 15 days. When a voting period opens, email notification is sent to NQF member organizations. Voting information also is made available on the NQF website. Each NQF member organization may cast one vote in favor of or against approval of a Standing Committee's recommendations. A member organization may also abstain from voting on a particular consensus development project.

All measures that are recommended by the Standing Committee or for which the Committee has not reached consensus, along with the results of member voting, will proceed to the next step in the CDP: review and recommendation by the CSAC. In rare instances, CSAC may request a second round of member voting.

- Consensus via NQF member voting has <u>not</u> been reached if the vote margin on any major criterion or overall is between 40%-60% in favor of endorsement.
    - When the NQF membership has not reached consensus for a measure, NQF will request that each council review the measure, via email or conference call, and provide input on the council perspective for the measure to the CSAC during their review.

Consensus Standards Approval Committee (CSAC)

The CSAC holds three in-person meetings annually and convenes monthly by conference call. All meetings are open to NQF members and the public and audience members have the opportunity to comment on the measures under consideration. Measure developers are expected to attend the call, which is generally one to two hours, and answer any questions from members of the CSAC. Information about each CSAC meeting is also available on the NQF website, including the meeting's agenda and materials and the physical location or dial-in information.

During its meeting, the CSAC reviews the recommendations of the Standing Committee, the public and member comments and the responses, and the results of NQF Member voting. After detailed review of a measure, the CSAC determines if consensus has been reached across the various NQF Member Councils.

CSAC Criteria for Decision-making

To ensure a consistent approach to endorsement decisions, the CSAC identified the following criteria to guide its decision-making. The CSAC's rationale for not endorsing a measure that had been recommended by a Standing Committee and approved by the membership will be documented and communicated to the public.

- Strategic importance of the measure. The CSAC will consider the value added of a measure, such as the strategic importance to measure and report on a measure, and assess whether a measure would add significant value to the overall NQF portfolio.

- Cross-cutting issues concerning measure properties. The CSAC will consider issues such as harmonization with other applicable measures in the NQF portfolio or risk adjustment methodology.

- Adequate consensus across stakeholders. The CSAC will consider concerns raised by councils and may conclude that additional efforts should be made to address these concerns before making an endorsement decision on the measure.

- Consensus development process concerns. The CSAC will consider process concerns raised during the CDP, such as insufficient attention to member comment or issues raised about committee composition.

CSAC Voting

- 60% approval for endorsement of a measure by voting CSAC members is required to recommend a measure for endorsement. A measure is not recommended for endorsement when the vote margin is less than 40% of voting CSAC members in favor of endorsement. The CSAC has not reached consensus if the vote margin on any major criterion or overall is between 40%-60% in favor of endorsement.

    - When the CSAC has not reached consensus, NQF will request that each council review the measure, via email or conference call, and provide input on the council perspective for the measure to the CSAC during their next meeting. After CSAC reviews the council input, they will re-vote on the measure.

        - After the re-vote, all measures that are recommended (>60% in favor of endorsement) by the CSAC will be forwarded to the NQF Board of Directors for ratification.

Following the call or meeting, all of the CSAC's decisions regarding a measure or measures are posted on the NQF website. In addition, all of the CSAC's recommendations are forwarded to the NQF Board of Directors for ratification within 2-3 weeks.

DEVELOPER REQUESTS FOR RECONSIDERATION OF A MEASURE NOT RECOMMENDED

If unsatisfied with the Standing committee reconsideration, a request for reconsideration may be made to the CSAC co-chairs.

- All requests related to the criteria must first go to the Standing committee as identified above.
- Written request must be submitted to the CSAC after the Standing committee reconsideration and no less than two weeks prior to the next scheduled CSAC meeting or conference call.
- The written request must cite the specific evaluation criteria or subcriteria that the developer thinks was not applied properly to the specific information as submitted and evaluated by the Standing committee.
- CSAC Co-Chairs will make a decision with optional input from the entire CSAC.

If the request for reconsideration is based on a question of whether the CDP was followed, developers may send a written request for reconsideration to the CSAC citing the issues with a specific CDP process step, how it was not followed properly, and how it resulted in the specific measure not being recommended.

- Written request must be submitted to the CSAC no less than two weeks prior to the next scheduled CSAC meeting or conference call.
- CSAC Co-Chairs will make a decision with optional input from the entire CSAC.

PROCESS FOR CSAC REVIEW

- Staff and Standing committee co-chairs compile all information for review by the CSAC co-chairs
- The options for the CSAC co-chairs include:
    - uphold the Standing committee final recommendation if the criteria were applied appropriately and process followed; or
    - ask for input from the CSAC, particularly if co-chairs think there is merit to the assertion of inappropriate application of the criteria or not following the CDP;
    - request additional expert input;
- if a breach in the CDP was identified, determine if it adversely affected the outcome for the specific measure;
    - if the criteria were not applied properly, provide explicit explanation and clarification to the Standing committee and ask them to re-evaluate the measure using the clarified guidance.

NQF BOARD OF DIRECTORS GRANTS ENDORSEMENT

CSAC decisions regarding consensus standards are submitted to the Board of Directors. The Board can affirm or deny a CSAC decision. All consensus standards that are recommended must be ratified by the Board for endorsement.

APPEALS

After a consensus standard has been formally endorsed by NQF, any interested party may file an **appeal of the endorsement decision** with the NQF Board of Directors. An appeal may only be filed in response to NQF endorsement of a candidate standard or set of standards; that is, an interested party *may not* file an appeal regarding the decision to not endorse a candidate standard. An interested party may file a concern about any measure (whether endorsed or not endorsed) in the NQF consensus development process and this concern will be reviewed by the CSAC.

An appeal of an endorsed measure must be filed within 30 days of the endorsement decision by going to the project webpage or the searchable list of all NQF-endorsed® national voluntary consensus standards. For an appeal to be considered by NQF, the appeal must include written evidence that the appellant's interests are directly and materially affected by the measure recently endorsed by NQF, and that NQF's endorsement of this measure has had, or will have, an adverse effect on those interests. All appeals are published on the NQF website.

Appeals are compiled and the CSAC reviews them and evaluates whether the concern raised is relevant and should warrant consideration of overturning the endorsement decision. After discussions, the CSAC will make a recommendation to the NQF Board of Directors regarding the appeal. The Board of Directors will take action on an appeal within seven calendar days of its consultation with the CSAC. The NQF Board of Directors' decision on an appeal of endorsement will be publicly available on NQF's website.

Project staff will notify developers when the appeals period will open and close, and at the close of the appeals period, staff will notify developers if any appeals were submitted on their measure(s). If an appeal was submitted, staff may request that developers (if necessary) provide a written response to the issues outlined in the letter of appeal. The letter of appeal will be discussed at the next CSAC in- person meeting or conference call. CSAC will review and discuss the letter of appeal and the developer's written response. The appellant will be asked to speak to their concerns and the developer will be provided an opportunity to respond. The developer will be asked to attend the CSAC call (~1-2 hours) and to answer any questions from CSAC. Following the CSAC call, staff will notify the developer of CSAC's recommendation to the NQF Board of Directors and will notify the developer of the Board decision on the appeal.

VI. Measure Evaluation Criteria

For details on NQF's measure evaluation criteria and guidance, please refer to the following resources:

- NQF's Measure Evaluation Criteria and
- NQF's Measure Evaluation and Criteria and Guidance on Evaluation.

Overview of NQF's evaluation criteria

Before being granted NQF endorsement, candidate performance measures must be evaluated against NQF's measure evaluation criteria. These criteria—which reflect desirable characteristics of performance measures—are used to determine the suitability of measures for use in both internal quality improvement efforts and in accountability applications. Currently, NQF has established five major evaluation criteria (see listing below). Subcriteria under each of the five major criteria have been formulated to help determine the extent to which the major criteria have been met. For example, the evidence, performance gap, and high priority subcriteria help to answer the question about whether and how a measure is important to measure and report. Most of these criteria and subcriteria apply to all types of measures, but a few are relevant to a specific type of measure and are noted as such.

Measure Evaluation Criteria *(abbreviated)*

1. **Importance to measure and report** (must-pass)
   - 1a. Evidence to Support the Measure Focus (must-pass)
   - 1b. Performance Gap, including disparities (must-pass)
   - 1c. High Priority (must-pass)
   - 1d. For composite measures: quality construct and rationale (must-pass)
2. **Scientific acceptability of measure properties** (must-pass)
   - 2a. Reliability [includes additional subcriteria] (must-pass)
   - 2b. Validity [includes additional subcriteria] (must-pass)
   - 2c. Disparities (addressed in 1b)
   - 2d. For composite measures: empirical analysis supporting composite construction (must-pass)
3. **Feasibility**
   - 3a. Required data elements routinely generated and used during care delivery
   - 3b. Availability in electronic health records or other electronic sources OR a credible, near-term path to electronic collection is specified
   - 3c. Data collection strategy can be implemented
4. **Usability and Use**
   - 4a. Accountability and Transparency
   - 4b. Improvement
   - 4c. The benefits to patients outweigh evidence of unintended negative consequences to patients
5. **Comparison to Related or Competing Measures**
   - 5a. Measure specifications are harmonized OR differences are justified
   - 5b. Superior measure is identified OR multiple measures are justified

The ordering of the criteria and subcriteria is deliberate, as is the designation of some criteria and subcriteria as "must-pass". NQF endorsement is intended to identify those performance measures that are most likely to facilitate achievement of high quality and efficient healthcare for patients. Thus, the first criterion—Importance to Measure and Report—reflects the goal of measuring those aspects with greatest potential of driving improvements. Specifically, measures that are Important to Measure and Report are evidence-based, reflect variation in performance, overall less-than-optimal performance, or disparities, and address a specific national health goal or priority or a high-impact aspect of healthcare. This criterion allows for a distinction between things that are important to do in clinical practice versus those that rise to the level of importance required for a national performance measure. NQF considers the Importance to Measure and Report criterion and its associated subcriteria as paramount: not only is importance the first criterion considered in the evaluation process, but it and all three subcriteria are must-pass criteria. That is,

if a measure does not meet one of the subcriteria under Importance to Measure and Report, it will not "pass" Importance and will not be endorsed. Procedures for voting to determine the Committee's evaluation of whether criteria are met are described elsewhere in this document.

Once the Standing Committee agrees that a measure is important to measure and report, it will then consider the scientific properties of the measure. The second evaluation criterion—Scientific Acceptability of Measure Properties—reflects NQF's view that performance measures must demonstrate sound measurement science—that is, they must be both reliable and valid. Measures that are reliable and valid enable users to make correct conclusions about the quality of care that is provided. Thus, both the reliability and validity subcriteria under the Scientific Acceptability criterion are must-pass subcriteria; if both of these are not met, then the measure will not be endorsed.

Once the Standing Committee agrees that a measure is scientifically acceptable (i.e., reliable and valid), it will then consider the feasibility of the measure. The Feasibility criterion reflects the extent to which the data required to compute a measure are readily available and retrievable without undue burden and the ease of implementation for performance measurement. The goal underlying this criterion is to endorse measures that cause as little burden as possible in terms of data collection and measure implementation. For example, the most feasible measures are those that use data from activities that are performed as part of the care delivery process and do not require separate additional or burdensome data collection and retrieval processes (e.g., data elements are stored in an electronic format such as an EHR). The Feasibility criterion is not considered must-pass. Assuming that a measure meets all the subcriteria for Importance and Scientific Acceptability, feasibility generally should not be the only reason that a measure would fail endorsement. In fact, feasibility may improve with broader implementation and ways to improve feasibility should be sought for important and sound performance measures.

The fourth criterion is that of Usability and Use. As noted earlier, NQF-endorsed measures are considered suitable for both accountability and quality improvement purposes and the expectation is that endorsed measures not only will be used, but also ultimately will lead to improved patient outcomes. Because it takes time for newly-developed measures to be selected for—and then implemented—in various programs, the Usability and Use criterion is not designated as must-pass for initial endorsement, although it becomes more critical when evaluating measures for continued endorsement.

Finally, if the Standing Committee agrees that a measure has met the first four NQF evaluation criteria, the Committee also will evaluate that measure in relation to measures that are similar. The current performance measure landscape contains an abundance of measures, including some that could be considered duplicative or overlapping and others that measure similar but somewhat different activities and/or patient populations. Such duplicative measures and/or those with similar but not identical specifications may increase data collection burden and/or create confusion or inaccuracy in interpreting performance results for those who implement and use those measures. The Comparison to Related or Competing Measures criterion requires a careful consideration of such similar measures, with the goal of endorsing only the best measures—or, if there isn't a "best" measure, endorsing measures that are consistent to the extent possible.

Rating scales

Usually the evaluation of a measure isn't a straightforward yes/no, all-or-nothing determination. Instead, measures typically meet the criteria to a greater or lesser extent. This is why NQF selects Standing Committee members who, collectively, have a wide variety of expertise and experience in a particular clinical area, in measurement, in using performance data, or in some other aspect of the quality enterprise.

To facilitate measure evaluation, NQF has developed several rating scales and algorithms to use when evaluating the criteria. For some criteria or subcriteria, a generic rating scale will suffice; for others, more specific rating algorithms have been developed. For the most part, however, all rating scales use the same four categories (high, moderate, low, and insufficient). Most criteria and subcriteria require a high or moderate rating from the Committee to "pass". Criteria rated with low or insufficient ratings generally do not pass—although these ratings reflect different underlying reasons for failure to pass. For example, a low rating generally means the evidence/information submitted actually demonstrates that a criterion has not been met; in contrast, a rating of insufficient means either that the information submitted is not adequate for a definitive answer or that the submission was incomplete or deficient in presenting existing evidence/information.

Evaluating new versus previously-endorsed measures

All measures—both new and previously-endorsed measures that are undergoing endorsement maintenance—are expected to meet the current criteria and guidance. However, the criteria and subcriteria differ somewhat depending on whether the measure has been previously endorsed. For example, by the time of measure maintenance, NQF expects measures to be in use—and therefore developers should submit data from implementation of the measure as specified to demonstrate performance gap (rather than using data from the literature). Similarly, experience with use—including any problems with implementation or unintended consequences—and data showing improvement on the performance measure should be included under the Usability and Use criterion. Also, by the time of measure maintenance (and use of the measure), there is an expectation that reliability and validity testing achieve a high rating, which requires testing at the performance measure score level (although testing for data elements only is acceptable at initial endorsement).

Closer look at NQF's evaluation criteria and subcriteria

This section is meant to provide a more detailed explanation of NQF's evaluation criteria and subcriteria by presenting, for each, some contextual information, key points for measure evaluation, and directions for finding relevant examples (when appropriate) in our companion document entitled What Good Looks Like. Additional detail regarding NQF's evaluation criteria and guidance can be found in various reports available on the measure evaluation criteria web page here (see links on the right-hand side of the webpage).

Criterion #1: Importance to Measure and Report

The criterion is meant to reflect the extent to which the specific measure focus—the activity or condition being measured—is evidence-based, important for making significant gains in health care quality, and improving health outcomes for a specific high-impact aspect of healthcare where there is variation in or overall less-than-optimal performance. The purpose of this criterion is to help focus measurement efforts on those things that are most likely to drive improvement in healthcare quality. It takes a lot of resources—in time, dollars, opportunity costs, etc.— to collect and transmit data, publish performance scores, and do

other activities within the improvement enterprise—and these limited resources should be expended on high-leverage activities. NQF recognizes that there are many things that are important to do in clinical practice, yet not all of these things necessarily rise to the level of importance required for endorsement by NQF as a national consensus standard for measuring performance.

NQF has a hierarchical preference for performance measures of health outcomes (including patient-reported outcomes) as follows:

- Outcomes linked to evidence-based processes/structures
- Outcomes of substantial importance with plausible process/structure relationships
- Intermediate outcomes that are most closely linked to outcomes
- Processes/structures that are most closely linked to outcomes

NQF's prefers outcome measures because:

- outcomes (e.g., improved function, survival, or relief from symptoms) are the reasons patients seek care and why providers deliver care;
- outcomes are of interest to purchasers and policymakers;
- outcomes are integrative, reflecting the result of all care provided over a particular time period (e.g., an episode of care);
- measuring performance on outcomes encourages a "systems approach" to providing and improving care; and
- measuring outcomes encourages innovation in identifying ways to impact or improve outcomes that might have previously been considered not modifiable (e.g., rate of central line infection).

Notwithstanding NQF's preference for outcome measures, there is also a need for other types of quality measures. Although there are countless intermediate outcomes, processes of care, and structural characteristics that influence health outcomes, NQF prefers measures of those that are the most closely linked (i.e., are most proximal) to desired outcomes.

KEY POINTS

- Limited resources are available for collecting data, measuring performance, and reporting performance results; NQF endorsement sets in motion an infrastructure that requires resources to accomplish these activities
- NQF endorsement of a measure as a national consensus standard requires a "higher bar" for importance than other measures that may be appropriate for use in internal quality improvement initiatives
- NQF has a hierarchical preference for outcome measures (including patient-reported outcomes), followed by intermediate clinical outcomes, then by process or structural measures that are proximal to desired outcomes
- All three subcriteria under Importance to Measure and Report are "must-pass"; therefore, each must be met in order to be recommended for endorsement

Subcriterion 1a: Evidence

This subcriterion is meant to address the question of whether there is an adequate level of empirical evidence to support a measure for use as a national consensus standard. The assumption underlying this subcriterion is that use of limited resources for measuring and reporting a measure is justified only if there is unambiguous evidence that it can facilitate gains in quality and health. For most healthcare quality measures, the evidence will be that of clinical effectiveness and a link to desired health outcomes (e.g., improved clinical outcomes, functional status, or quality of life; decreased mortality; etc.). The strength of such evidence is related to its ***quantity, quality,*** and ***consistency*** from the relevant body of evidence.

For process measures, structural measures, and measures of intermediate outcomes, the quantity, quality, and consistency of the body of evidence underlying the measure should demonstrate that the measure focuses on those aspects of care known to influence desired patient outcomes (i.e., those with the most direct evidence of a strong relationship to the desired outcome). For example, evidence about effective medication to control blood pressure is direct evidence for the medication but only indirect evidence for the frequency of assessing blood pressure; assessing blood pressure, although necessary, is not sufficient for achieving control.

Evidence refers to empirical studies, but is not limited to randomized controlled trials. The preferred sources of evidence are systematic reviews and grading of a body of evidence that are conducted by independent organizations (e.g., USPSTF, Cochrane Collaboration, etc.). Because not all healthcare is evidence-based, NQF will allow—under certain circumstances—an exception to the evidence subcriterion; however, granting of such exceptions should not be considered routine.

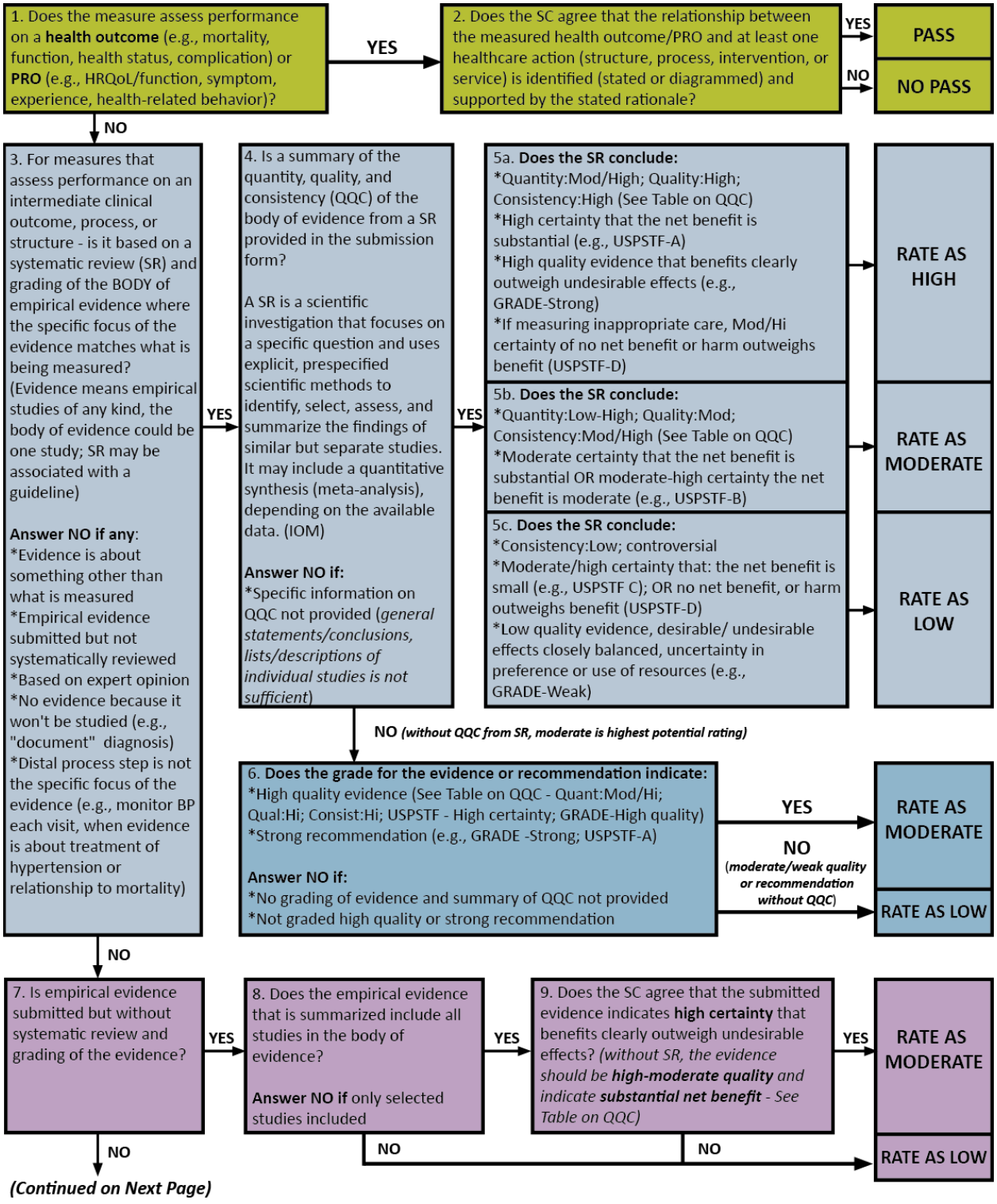
For health outcome measures and patient-reported outcome performance measures, NQF currently does not require a summary of a systematic review of the empirical evidence that links the outcomes to certain processes and/or structures of care because there are myriad processes and structures that may influence health outcomes. However, NQF does require that developers of these types of measures articulate a rationale (which often includes evidence) for how the outcome is influenced by healthcare processes or structures. The evidence subcriterion is not applicable to resource use measures.

Guidance for evaluating the clinical evidence is provided in Algorithm #1.

KEY POINTS ON EVALUATING EVIDENCE

- The evaluation of the evidence subcriterion depends on the type of measure under consideration
- Evidence should be presented about the relevant body of evidence—not selected individual studies
- Ideally, measure developers will summarize a systematic review of the evidence that has been assembled, reviewed, and graded by others
- Expert opinion is not considered to be empirical evidence, but evidence is not limited to randomized controlled trials
- Measures with inconsistent or conflicting evidence should not pass the evidence subcriterion
- When evaluating the quality of the evidence, consider the following:
  - The study design itself (e.g., RCT, non-RCT) or flaws in the design or conduct of the study (e.g., lack of allocation concealment or blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; failure to report important outcomes)
  - The directness/indirectness of the evidence to the measure as specified (e.g., regarding the population, intervention, comparators, and/or outcomes)
  - Imprecision in study results (i.e., wide confidence intervals due to few patients or events)
- Under limited circumstances, an exception to the evidence subcriterion may be invoked and evaluated according to the evidence algorithm

Algorithm #1. Guidance for Evaluating the Clinical Evidence



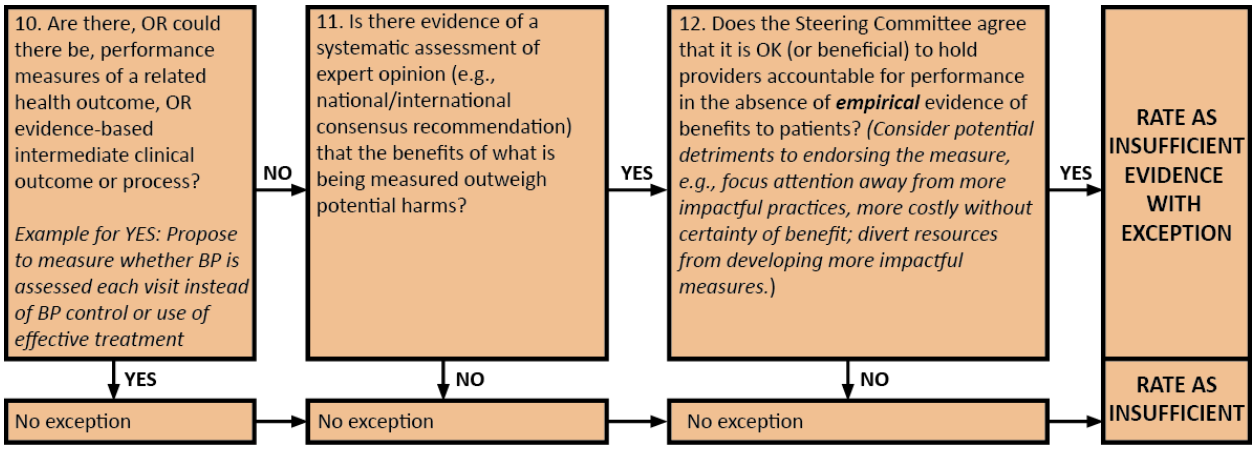


Table 1: Evaluation of Quantity, Quality, and Consistency of Body of Evidence for Structure, Process, and Intermediate Outcome Measures (to be used with Algorithm #1)

DEFINITION /RATING	**QUANTITY** OF BODY OF EVIDENCE	**QUALITY** OF BODY OF EVIDENCE	**CONSISTENCY** OF RESULTS OF BODY OF EVIDENCE
Definition	Total number of studies (not articles or papers)	Certainty or confidence in the estimates of benefits and harms to patients across studies in the body of evidence related to study factors[a] including: study design or flaws; directness/indirectness to the specific measure (regarding the population, intervention, comparators, outcomes); imprecision (wide confidence intervals due to few patients or events)	Stability in both the direction and magnitude of clinically/practically meaningful benefits and harms to patients (benefit over harms) across studies in the body of evidence
High	5+ studies[b]	Randomized controlled trials (RCTs) providing direct evidence for the specific measure focus, with adequate size to obtain precise estimates of effect, and without serious flaws that introduce bias	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction and similar in magnitude across the preponderance of studies in the body of evidence

DEFINITION /RATING	**QUANTITY** OF BODY OF EVIDENCE	**QUALITY** OF BODY OF EVIDENCE	**CONSISTENCY** OF RESULTS OF BODY OF EVIDENCE
Moderate	2-4 studies[b]	• Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect OR • RCTs without serious flaws that introduce bias, but with either indirect evidence or imprecise estimate of effect	Estimates of clinically/practically meaningful benefits and harms to patients are consistent in direction across the preponderance of studies in the body of evidence, but may differ in magnitude If only one study, then the estimate of benefits greatly outweighs the estimate of potential harms to patients (one study cannot achieve high consistency rating)
Low	1 study[b]	• RCTs with flaws that introduce bias OR • Non-RCTs with small or imprecise estimate of effect, or without control for confounders that could account for other plausible explanations	• Estimates of clinically/practically meaningful benefits and harms to patients differ in both direction and magnitude across the preponderance of studies in the body of evidence OR • wide confidence intervals prevent estimating net benefit If only one study, then estimate of benefits do not greatly outweigh harms to patients
Insufficient to Evaluate	• No empirical evidence OR • Only selected studies from a larger body of evidence	• No empirical evidence OR • Only selected studies from a larger body of evidence	No assessment of magnitude and direction of benefits and harms to patients

[a] *Study designs* that affect certainty of confidence in estimates of effect include: randomized controlled trials (RCTs), which control for both observed and unobserved confounders, and non-RCTs (observational studies) with various levels of control for confounders. *Study flaws* that may bias estimates of effect include: lack of allocation concealment; lack of blinding; large losses to follow-up; failure to adhere to intention to treat analysis; stopping early for benefit; and failure to report important outcomes. *Imprecision* with wide confidence intervals around estimates of effects can occur in studies involving few patients and few events. *Indirectness* of evidence includes: indirect comparisons (e.g., two drugs compared to placebos rather than head-to head); and differences between the population, intervention, comparator interventions, and outcome of interest and those included in the relevant studies.

[b]The suggested number of studies for rating levels of quantity is considered a general guideline.

Example

What Good Looks Like: Process #1 (pp. 4-9); Process #2 (pp. 4-10); Outcome (pp. 3-6)

Subcriterion 1b: Performance Gap

This subcriterion is meant to address the question of whether there is actually a quality problem that is addressed by a particular measure. Again, because the measurement enterprise is resource intensive, NQF's position is to endorse measures that address areas of known gaps in performance (i.e., those for which there is actually opportunity for improvement). Opportunity for improvement can be demonstrated via data that indicate overall poor performance (in the activity or outcome targeted by the measure), substantial variation in performance across providers, or variation in performance for certain subpopulations (i.e., disparities in care).

Occasionally, measures that are being evaluated for continued endorsement may reflect a high level of performance across all providers and for all population subgroups (that is, they may be "topped out"). Such measures typically would not meet the performance gap subcriterion and thus would not be granted continued endorsement. However, for some such measures, the impact of loss of endorsement may be serious and the Standing Committee could consider recommending those measures for Reserve Status if continued monitoring is needed to ensure that performance does not decline. Use of the Reserve Status should be an exception—not the rule. Further, it can be applied only to highly credible, reliable, and valid measures that have high levels of performance due to quality improvement actions (e.g., not due to documentation practices only) .

The rating scale used for evaluating performance gap is provided in Table 2.

Table 2: Generic Scale for Rating Subcriteria 1b, 1c, 1d and Criteria 3 and 4

RATING	DEFINITION
High	Based on the information submitted, there is high confidence (or certainty) that the criterion is met
Moderate	Based on the information submitted, there is moderate confidence (or certainty) that the criterion is met
Low	Based on the information submitted, there is low confidence (or certainty) that the criterion is met
Insufficient	There is insufficient information submitted to evaluate whether the criterion is met (e.g., blank, incomplete, or not relevant, responsive, or specific to the particular question)

KEY POINTS FOR OPPORTUNITY FOR IMPROVEMENT

- Ideally, demonstration of opportunity for improvement for a particular measure should be based on data for that particular measure as specified; however, relevant data from the literature also may be used, especially for initial endorsement
- When evaluating whether there is opportunity for improvement, consider:
  - The distribution of performance scores
  - The number and representativeness of the entities included in the measure performance data
  - The size of the population at risk, effectiveness of an intervention, likely occurrence of an outcome, and consequences of the quality problem
  - Data on disparities

Example

Currently under development.

Subcriterion 1c: High priority

This subcriterion is meant to address the question of whether the focus of a particular measure addresses a specific national health goal or priority and/or a high-impact aspect of healthcare. For example, the property of "high priority" is demonstrated when a measure is aligned with one of the National Quality Strategy priorities or with a specific national health goal (e.g., reducing hospital readmissions). Alternatively, a measure can be considered as addressing a high-priority aspect of healthcare if epidemiologic or resource use data demonstrates that the measure can affect large numbers of patients and/or has a substantial impact for a smaller population, if the associated condition is a leading cause of morbidity/mortality, and/or if the associated condition results in high resource use (current and/or future), high illness severity, or if the consequences of poor quality would severely impact patient or societal health. Most performance measures can be somehow associated with the broad NQS priorities, and developers are asked to provide epidemiologic or resource use data.

The rating scale used for evaluating performance gap is provided in Table 2.

KEY POINTS FOR PRIORITY

- Epidemiologic or resource use data to demonstrate high priority should be included (e.g., number of persons or percentages affected, dollar amounts, etc.), not just statements or conclusions

Example

Currently under development.

Subcriterion 1d: Quality construct and rationale (relevant to composite performance measures only)

A composite performance measure is a combination of two or more component measures, each of which individually reflects quality of care, into a single performance measure with a single score (the types of measures that will and will not be considered composite performance measures for purposes of NQF measure submission, evaluation, and endorsement are listed on page 5 of the **Composite Performance Measure Evaluation Guidance report**). The first step in developing a composite performance measure should be to articulate a coherent quality construct and rationale to guide construction of the composite. Once this is determined, the developer should select which component measures will be included in the composite measure and determine how those components will be combined.

This subcriterion allows measure developers to "tell the story" behind their composite performance measure. Specifically, developers are asked to describe the quality construct, which should include the following:

- overall area of quality (e.g., quality of CABG surgery);
- component measures that are included in the composite performance measure (e.g., pre-operative beta blockade; CABG using internal mammary artery; CABG risk-adjusted operative mortality);
- conceptual relationships between each component and the overall composite (e.g., components cause or define quality, components are caused by or reflect quality); and
- relationships among the component measures (e.g., whether they are correlated or not, processes that are expected to lead to better outcomes).

They should also describe the rationale underlying the composite performance measure, including a discussion of how the composite performance measure provides added value over and above what is provided by the component measures individually. Finally, they should describe how their method for combining the component measures "fits" with the quality construct and rationale that they have articulated.

The rating scale used for evaluating the quality construct and rationale is provided in Table 2.

KEY POINT ON QUALITY CONSTRUCT FOR COMPOSITE MEASURES

- This subcriterion allows developers to "tell their story" of how they conceptualized and then built the composite performance measure

Criterion #2: Scientific Acceptability of Measure Properties

The criterion is meant to reflect the extent to which the measure, as specified, produces consistent and credible results about the quality of care. The focus of this criterion is measurement science—not clinical science (which is the focus of the evidence subcriterion under Importance to Measure and Report).

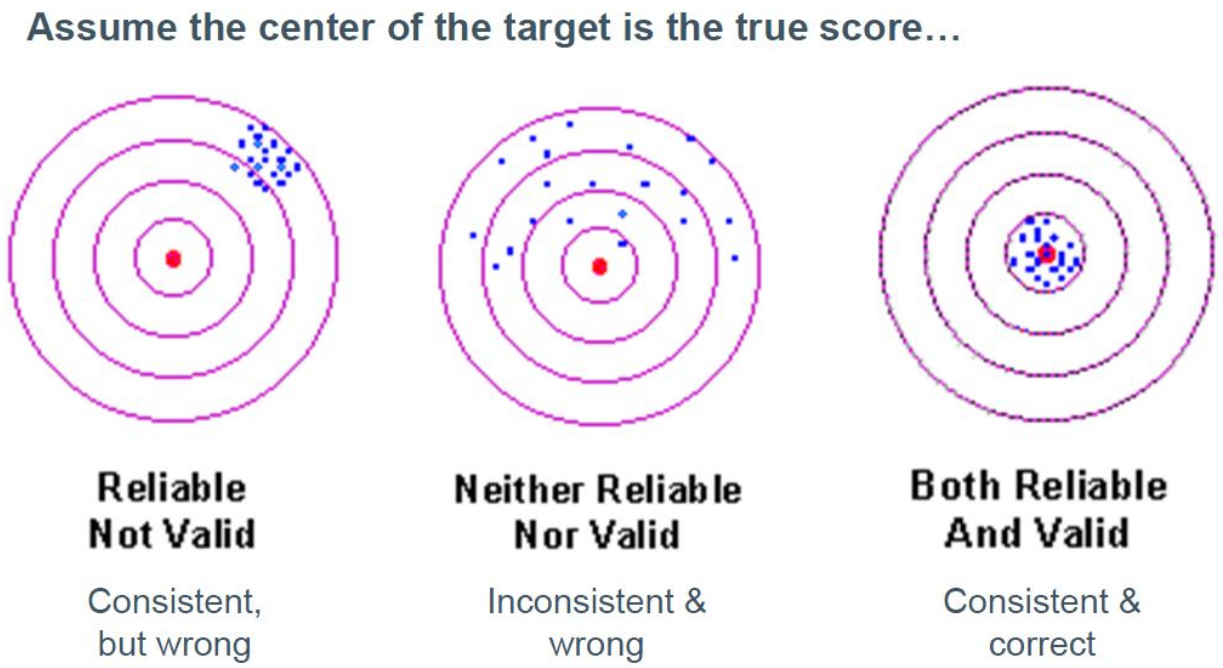
Specifically, this criterion addresses the basic measurement principles of *reliability* and *validity.* Consideration of reliability and validity can help to address the following questions:

- Are the specifications clear so that everyone will calculate the measure in the same way?
- Are the specifications consistent with the evidence?
- Is the variation between providers primarily due to real differences? Or is it because there is a lot of "noise" in the measurement?
- Is the measure actually measuring what it is intended to measure (i.e., quality of care)?
- Do the results of the measurement allow for correct conclusions about quality of care?

Use of measures that are unreliable or invalid could result in inconsistent measurement, inaccurate measurement, measurement that cannot differentiate providers, and/or measurement that leads to wrong conclusions about the quality of care that is provided. The consequences of using unreliable or invalid measures can be considerable (e.g., waste of resources used in data collection, reporting, and reacting to results; misinformation, misdirection, or even unintended harmful consequences for patients). Ultimately, the use of unreliable or invalid measures will undermine confidence in measures among providers and consumers of healthcare.

Figure 1 (adapted from figure at http://www.socialresearchmethods.net/kb/relandval.php) illustrates the concepts of reliability and validity of measurement. The center of the target is the concept that is being measured (e.g., percentage of facilities that provide aspirin to heart attack patients within 24 hours of arrival). Each dot on the target represents a measurement. In the first target, all of the measurements are quite similar (and consistent), but they don't do a very good job of hitting the target—this portrays a measure that is reliable, but not valid. In the second target, the measurements aren't very close to each other or to the center of the target—this portrays a measure that is neither reliable nor valid. In the third target, all of the measurements are close to each other and to the center of the target—this portrays a measure that is both valid and reliable. Note that in order to be valid, a measure must be reliable; however, reliability does not guarantee validity.

Figure 1. Schematic of reliability and Validity



Measure developers conduct empirical analyses—collectively referred to as *measure testing*—in order to demonstrate the reliability and validity of a measure. Various methods and statistics can be used to quantify reliability and validity, although some may be more appropriate than others. However, evaluating reliability and validity requires more than simply examining the results of measure testing: it also requires consideration of how the measure is constructed—are the specifications written so that the measure can be computed consistently and do those specifications conform to the evidence—and potential threats to reliability and validity. For example, vague or unclear specifications for a measure can result in random errors in data collection or scoring, which reduces reliability; inappropriate exclusion of a certain subpopulations from a measure can lead to incorrect conclusions about the quality of care that is provided, thus invalidating the measure.

Testing measures for reliability and validity—while necessary—does require resources. NQF criteria allow flexibility for measure developers to determine the most appropriate and efficient methods for testing. For example, developers can:

- conduct testing at either the data element level (using patient-level data) or at the performance measure score level (using data that have been aggregated across providers) for initial endorsement;
- conduct testing on samples of patients and providers;
- rely on existing evidence of reliability and/or validity if available for the specific measure and data elements (e.g., from the literature);
- "substitute" data element validity testing in place of data element reliability testing; and/or

- present evidence of the face validity of the performance measure score as an indicator of quality rather than conduct empirical validation (although the latter is preferred).

This flexibility can, however, make it more difficult for Standing Committees to evaluate the scientific merits of measures in a consistent manner. Therefore, NQF has developed algorithms to guide Standing Committee evaluation of measure reliability and validity (see Algorithms #2 and #3).

KEY POINTS ON RELIABILITY AND VALIDITY

- Scientifically acceptable measures must be both reliable and valid
- Empirical demonstration of reliability and validity is expected, although demonstration of face validity as an indicator of quality also is allowed
- NQF is not prescriptive about how empirical measure testing is done; similarly, NQF does not set minimum thresholds for reliability or validity
- Reliability and validity must be demonstrated for the measure as specified (including data source and level of analysis)
- NQF allows testing at either the data element level (using patient-level data) or at the performance measure score level (using data that have been aggregated across providers)
- When evaluating measure testing results, the method of testing, the data used for testing (often from a sample), and the results of the testing must be considered

Subcriterion 2a: Reliability

The ability to distinguish performance across providers is critical for measures that are used in accountability applications (e.g., certification, public reporting, payment incentives, etc.). In the field of quality performance measurement, reliability is a way of quantifying the chance error (or "noise") in a measure. All measures have some error—but when there is a lot of error in a measure, it can be difficult to know whether (or how much) variation in performance scores between providers is due to "real" differences between providers or to measurement error. Yet a performance measure is useful only if it can detect differences across those being measured (reliability), and when those differences represent differences in quality (validity) and not just differences due to chance. Because NQF endorsement implies suitability of a measure for use in both internal quality improvement efforts and in accountability applications, an evaluation of reliability is essential.

The foundation for a reliable measure starts with good specifications: definitions, codes, and instructions on how to calculate the measure. However, good specifications alone do not guarantee reliability—and therefore NQF's evaluation criteria require empirical testing of reliability. Developers can test reliability at the data element level, the performance measure score level, or both; note, however, that data element reliability testing is not required if data element validity has been demonstrated. Testing at the data element level addresses the *repeatability/reproducibility* of the patient-level data used in the measure; such testing should be done for all "critical" data elements (i.e., those needed to calculate the measure score), or, at a minimum, for the numerator, denominator, and exclusions. In contrast, testing at the performance measure score level addresses the *precision* of the measure; such testing uses data that have

been aggregated across providers. Developers also can choose from a variety of methods and statistics to test reliability. NQF is not prescriptive about the methods nor about the results; however, when evaluating the reliability of a measure, Standing Committees should consider the appropriateness of the method, the adequacy of the sample used in testing, and the results of the testing.

The additional subcriteria under subcriterion 2a (reliability) include:

2a1. Precise specifications, including exclusions

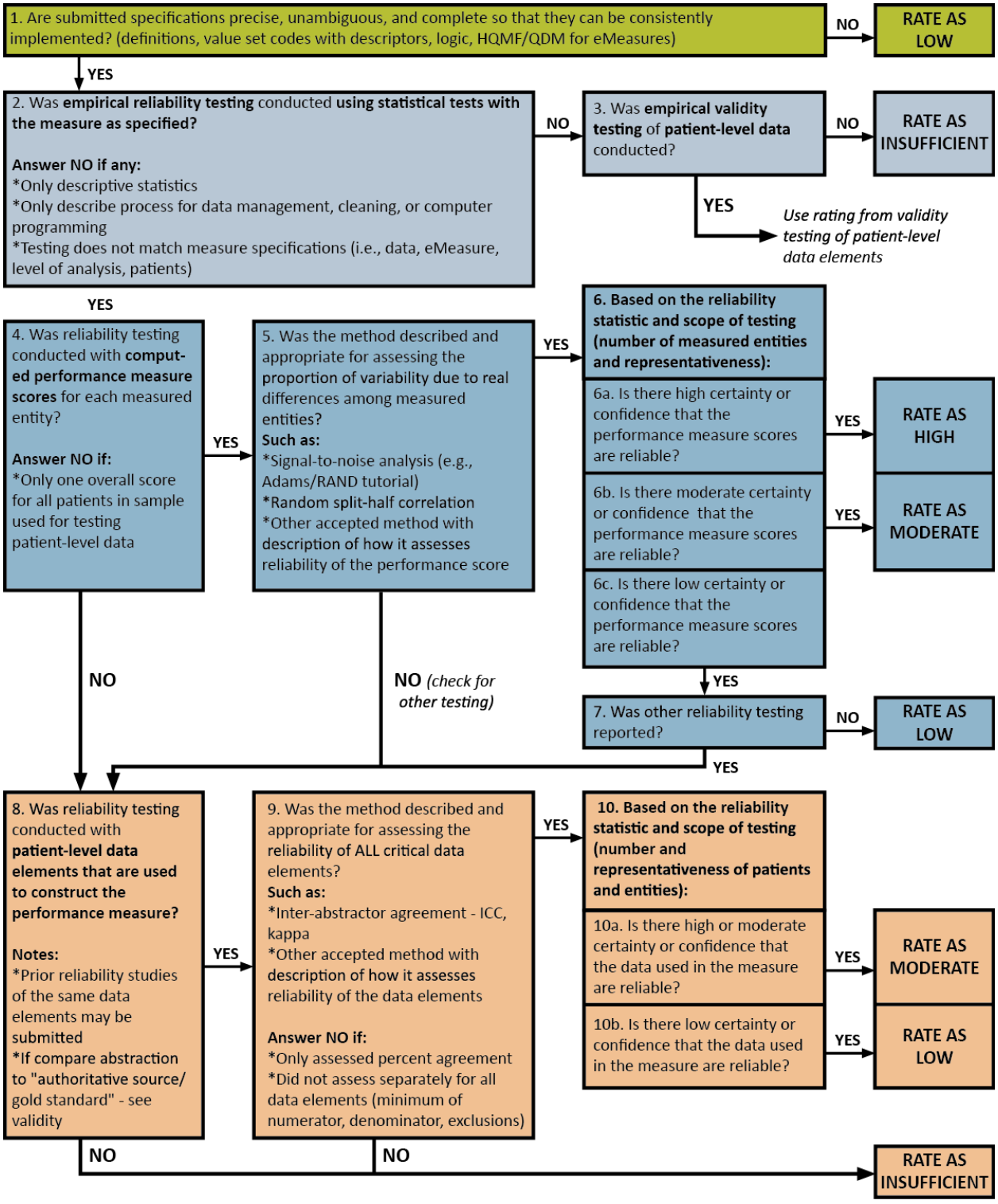
2a2. Reliability testing—data elements or measure score

Guidance for evaluating reliability is provided in Algorithm #2.

KEY POINTS ON RELIABILITY

- Reliability refers to the repeatability and precision of measurement
- Measurement precision reflects the ability to distinguish differences between providers that are due to quality of care rather than chance
- Precise specifications provide the foundation for achieving consistency in measurement
- Requirements for eMeasures specifications include use of:
  - the Health Quality Measures Format (HQMF)
  - the Quality Data Model (QDM)
  - value sets vetted through the National Library of Medicine's Value Set Authority Center (VSAC)
- Testing should be done for the measure as specified (including data source and level of analysis)
- Data element reliability
  - Addresses the repeatability/reproducibility of the data used in the measure
  - Uses patient-level data
  - Required for all critical data elements (i.e., those needed to calculate the measure score), or, at a minimum, for the numerator, denominator, and exclusions
  - Common method is inter-rater reliability (common statistics include kappa; intra-class correlation coefficient)
  - Not required if *data element validity* is demonstrated
- Performance measure score reliability
  - Addresses the precision of the measure
  - Uses data that have been aggregated across providers
  - Common method is signal-to-noise analysis
- When evaluating the testing of the measure, consider whether
  - an appropriate method was used
  - an adequate number of representative providers and patients were included

Algorithm #2. Guidance for Evaluating Reliability



Example

What Good Looks Like: Measure Testing (pp. 5-8)

Subcriterion 2b: Validity

The validity of a measure refers to the extent to which one can draw accurate conclusions about a particular attribute based on the results of that measure. In the context of quality performance measurement, a valid measure will allow one to make correct conclusions about the quality of care (e.g., a higher score on a quality measure reflects higher quality of care).

As with reliability, measure specifications are critical achieving measure validity: but when considering the validity, it is not precision of the specifications that is of interest, but rather, whether or not the specifications conform to the evidence underlying the measure. For example, if a measure of blood pressure control is specified, then the blood pressure threshold(s) used in the measure must conform to those indicated by the evidence (e.g., the level below which mortality and morbidity are reduced).

There are two general approaches for demonstrating validity: empirical testing or soliciting expert opinion. Face validity of a performance measure—the subjective determination by experts that, on the face of it, the measure appears to reflect quality of care—is the weakest demonstration of validity, but is accepted by NQF on initial endorsement. As with reliability testing, developers can choose from a variety of methods and statistics to test validity empirically.

Although there are various terms that sometimes are used to describe types of empirical validity testing, at its core, the validation process is one of assessing relationships. The developer should link the concept of interest (that is being measured) to some other concept(s) and articulate a hypothesis about the relationship between them. Usually many such linkages and hypotheses can be made—but both should be based on knowledge and understanding of the assumptions underlying the measure. Because the linkages and hypotheses are based on a theoretical understanding of the measure, developers should be able to explain the relationship(s) they expect to see (e.g., the magnitude or strength of the relationship and its direction, whether positive or negative). Developers will then test their hypotheses, and the results will provide information about the validity of the measure. For example, if the expected relationship is found, then it is likely that the hypothesis is sound and validity therefore has been demonstrated to some extent; conversely, if the expected relationship is not found, then either hypothesis itself or measure (or both) is at fault.

Developers can test validity at the data element level, the performance measure score level, or both. Testing at the data element level typically addresses the correctness of the patient-level data elements used in the measure, as compared to an authoritative source; such testing should be done for all "critical" data elements (i.e., those needed to calculate the measure score), or, at a minimum, for the numerator, denominator, and exclusions. In contrast, testing at the performance measure score level addresses the correctness of conclusions about quality that can be made based on the measure score; such testing uses data that have been aggregated across providers. Again, NQF is not prescriptive about the methods used in validity testing, nor about the results; however, when evaluating the validity of a measure, Standing Committees should consider whether the hypothesis is conceptually sound, the appropriateness of the testing method, the adequacy of the sample used in testing, and the results of the testing. Ideally, demonstration of validity should be accumulated over time, as additional testing is conducted using various methodologies and in various conditions.

Demonstration of validity also requires consideration of potential threats to validity (which can vary depending on the type of measure). Threats to validity may stem from other aspects of the measure specifications, including inappropriate exclusions, lack of appropriate risk adjustment or risk stratification for outcome and resource use measures, use of multiple data sources or methods that result in different scores and conclusions about quality. Other threats to validity may include systematic missing or "incorrect" data used in calculating the measure or unreliability of the measure itself. Most importantly, a measure may be invalid because the measurement has not correctly captured the concept of quality that it was intended to measure.

The additional subcriteria under subcriterion 2b (validity) include:

2b1. Specifications consistent with evidence
2b2. Validity testing
2b3. Justification of exclusions
2b4. Risk adjustment (for outcome and resource use measures)
2b5. Identification of differences in performance
2b6. Comparability of data sources/methods
2b7. Missing data (for eMeasures, composites, and PRO-PMs)

Note that some of these subcriteria may not be relevant for all measures.

Risk adjustment

Risk adjustment (also called case-mix adjustment) is the process of controlling for patient factors that are present at the start of care that could influence patient outcomes or resource use. The purpose of risk-adjustment is to "level the playing field" so that, when comparing providers, the differences in performance scores are due to differences in the quality of care provided rather than to differences in the patient groups (e.g., one provider's patients may be sicker than those of another provider).

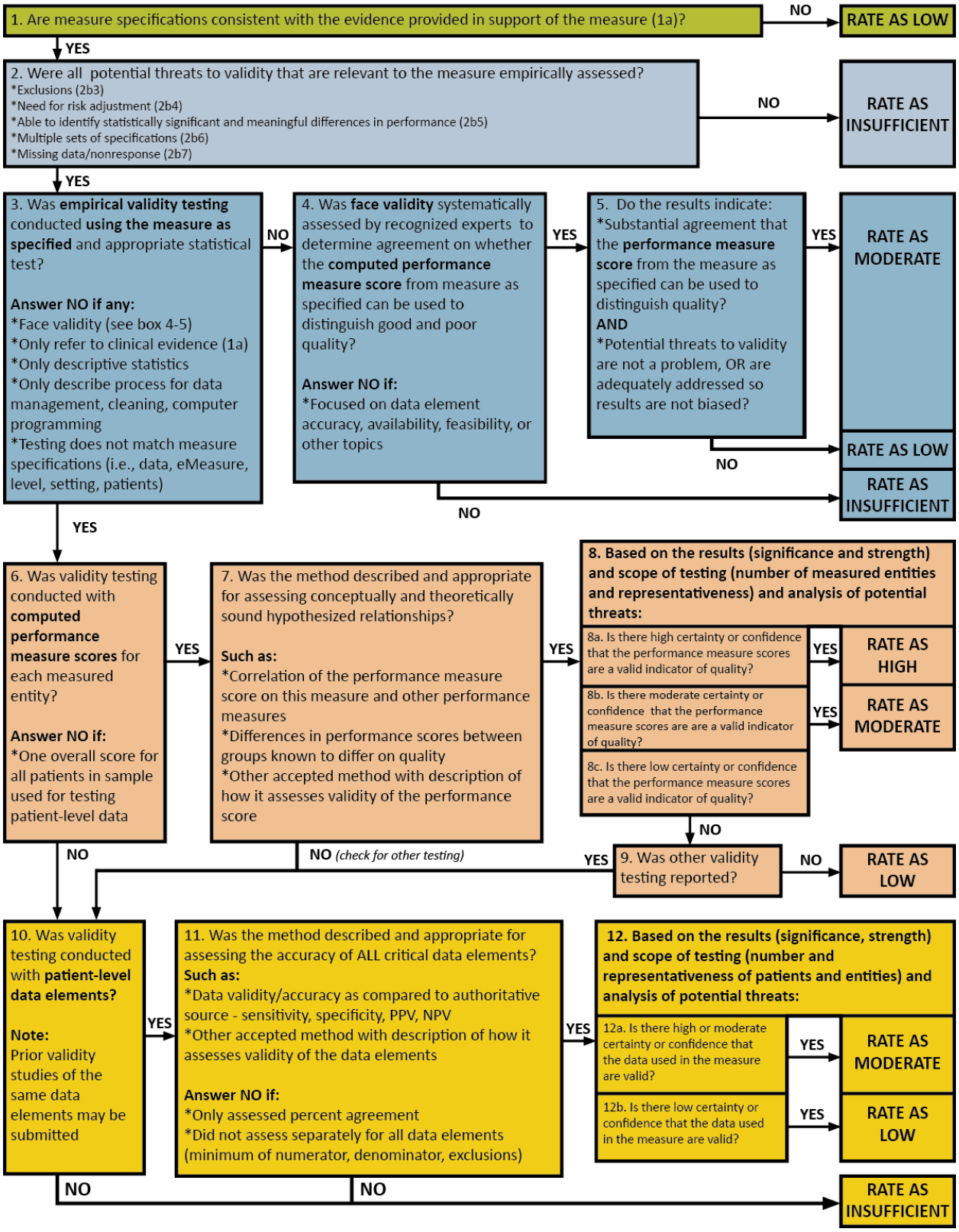
Factors used in risk adjustment should include patient-level factors that are associated with the outcome of interest but are not confounded with the quality of care that is provided. Thus, these factors should represent patient characteristics that are present at the start of care (e.g., severity of illness) and should not include structures/characteristics of organizations/clinicians associated with quality (e.g., experience, training, equipment). NQF's current guidance is that statistical risk adjustment models should not include factors associated with disparities (e.g., race, income) so as not to obscure any disparities in care; instead, measures should be stratified by such factors. Whether or not to modify this guidance (in some way) is the subject of a current NQF project; recommendations from this project are expected by June 2014.

Guidance for evaluating validity is provided in Algorithm #3.

KEY POINTS ON VALIDITY

- Validity refers to the *correctness* of measurement: that one is, in fact, measuring what he/she is intending to measure and that the results of the measurement allow one to make the right conclusions
- Specifications that are *consistent with evidence* provide the foundation for achieving measure validity
- Testing should be done for the measure as specified (including data source and level of analysis)
- Data element validity
  - Typically addresses the correctness of the data elements as compared to an authoritative source
  - Uses patient-level data
  - Must be done for all critical data elements (i.e., those needed to calculate the measure score)
  - Common method is analysis of agreement compared to an authoritative source (common statistics include sensitivity; specificity)
- Performance measure score validity
  - Addresses the correctness of conclusions about quality that can be made based on the measure scores
  - Uses data that have been aggregated across providers
  - Some typical analytical methods include:
    - Assessment of ability to predict or explain a score on some other theoretically related measure (e.g., scores on process performance measure predict scores on relevant outcome performance measure)
    - Correlation of the score with another related measure
    - Assessment of ability to distinguish between groups known to have higher and lower quality assessed by another valid method
- When evaluating empirical validity testing of the measure, consider whether
  - the hypothesis was conceptually sound
  - an appropriate method was used
  - an adequate number of representative providers and patients were included
  - the results of the testing were adequate (i.e., within acceptable norms)
  - potential threats to validity are adequately assessed and accounted for
- Face validity—the subjective determination that, on the face of it, a measure appears

Algorithm #3. Guidance for Evaluating Validity



Example

What Good Looks Like: Measure Testing (pp. 8-11)

Subcriterion 2c: Disparities

This subcriterion is now addressed under subcriterion 1b (performance gap), under Importance to Measure and Report.

Subcriterion 2d: Empirical analysis supporting composite construction (relevant to composite performance measures only)

While subcriterion 1d addresses the conceptual basis of the composite performance measure, this subcriterion allows developers to demonstrate—via *empirical* analyses—that the choices made regarding which components are included in the composite performance score and how those components are combined actually fits with their concept of quality. In reality, this subcriterion is an extension of the reliability and validity subcriteria; however, it is listed as a separate criterion to signify that it is specific to composite performance measures. As with reliability and validity, NQF is not prescriptive about the methods used in the analyses that address this subcriterion: in fact, the methods used should follow from the quality construct that is described in subcriterion 1d.

KEY POINTS FOR COMPOSITE MEASURES

- This subcriterion allows developers to demonstrate empirically that the choices about which component measures are included in the composite and how those components are combined is consistent with their stated quality construct
- If empirical analyses do not provide adequate results (or are not conducted), other justification must be provided (and accepted by the Standing Committee) in order to pass this subcriterion

Criterion #3: Feasibility

This criterion is intended to assess the extent to which the specifications—including measure logic—require data that are readily available or could be captured without undue burden and can be implemented for performance measurement. The first two subcriteria under Feasibility relate to the burden of data collection and the third subcriterion relates to ease of implementation.

The feasibility of eMeasures hinges on the data elements that are included in the measure and the logic that is used to compute the measure. Thus, for eMeasures, a summary of a feasibility assessment is required. Ideally, developers would utilize a standard scorecard to reflect this summary (see Table 6 in the NQF Measure Evaluation Criteria and Guidance document for an example of a data element feasibility scorecard). At a minimum, however, the summary would include a description of the assessment; feasibility scores for all data elements, along with explanatory notes for all data element components with a low feasibility rating; attestation that the measure logic can be executed; and a rationale and plan for addressing any feasibility concerns.

The rating scale used for evaluating Feasibility is provided in Table 2.

KEY POINTS ON FEASIBILITY

- The feasibility criterion is concerned with the burden of data collection and the ease of implementation of the measure
- When evaluating the feasibility of the measure, consider whether
    - the required data elements are routinely generated and used during care delivery
    - the required data elements available in electronic form (e.g., EHR or other electronic sources)
    - the data collection strategy is ready to be put into operational use
- A summary of a formal feasibility assessment should be provided for eMeasures
- Feasibility is not a must-pass criterion

Criterion #4: Usability and Use

This criterion is intended to assess the extent to which potential audiences (e.g., consumers, purchasers, providers, policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high-quality, efficient healthcare for individuals or populations. Note that NQF currently does not endorse measures that are intended only for use in internal quality improvement efforts; instead, there is an expectation that NQF-endorsed measures will be used both internally for improvement as well as externally for accountability.

Measures are not required to be in use at the time of initial endorsement (although a plan and timeline for implementation should be provided); however, measures ideally should be in use in accountability programs by the time of endorsement maintenance and be publicly reported within six years of initial endorsement. However, the Usability and Use criterion goes beyond simply requiring that measures be used: it also reflects the desire that measures are demonstrably useful for improvement. In addition, this criterion also reflects the need for consideration of unintended negative consequences of the measure to individuals or populations (if any). This consideration should not center on theoretical negative consequences but instead should be those that are supported by evidence (e.g., the nature of the unintended negative consequence, the affected party, the number of people affected, and the severity of the impact).

The rating scale used for evaluating Usability and Use is provided in Table 2.

KEY POINTS ON USE AND USABILITY

- Measures are not required to be in use at initial endorsement, but ideally should be used in at least one accountability application by the time of endorsement maintenance and be publicly reported within six years of initial endorsement
- If not in use at time of initial endorsement, a credible plan for use and credible rationale for improvement should be provided
- If a measure is not in use in an accountability application or in public reporting by the time of endorsement maintenance, the reasons should be articulated and a credible plan for implementation/public reporting should be provided
- By the time of endorsement maintenance, some evidence that the measure results in improvement in health and/or healthcare is required
- Evaluation of this criterion will include a consideration of unintended negative consequences
- When evaluating the use and usability of a measure, consider whether
    - it is used in at least one accountability application or is publicly reported
    - the performance results have been used to further the goal of high-quality, efficient healthcare
    - the benefits of the measure outweigh any potential unintended

Criterion #5: Related and Competing Measures

NQF endorses national standards—and this implies parsimony and standardization to the extent possible. Duplicative measures and/or those with similar but not identical specifications increases measurement burden can create confusion or inaccuracy in interpreting performance results, especially if such measures produce different results for the same provider. Therefore, if a measure has met all the previous NQF evaluation criteria, the Standing Committee will then evaluate that measure in relation to other competing or related measures. In this evaluation, the two primary considerations will be the evidence driving the differing measure specifications the applicability of the measure (ideally, measures should include as many relevant entities as possible, based on the evidence).

Competing measures are those measures that are intended to address the same measure focus and the same target population, while related measures are those intended to address the same measure focus or the same target population. Ideally, when evaluating competing measures, the Committee will be able to identify the superior measure(s)—in in which case, the Committee would recommend the superior measure as suitable for endorsement but would not recommend the competing measures. Similarly, when evaluating related measures, the Committee ideally will be able to make recommendations for harmonization (suggested alterations of related measures to make their specifications more similar). The dimensions of harmonization can include numerator, denominator, exclusions, calculation, and data source and collection instructions; however, the extent of harmonization depends on the relationship of the measures, the evidence for the specific measure focus, and differences in data sources. In some cases, there may be valid reasons to endorse competing measures or measures that are not harmonized to the

extent possible, and measure developers have the opportunity to justify this course of action for the Committee.

There is no rating scale for the evaluation of competing or related measures; instead, staff will guide the Committee through a discussion of relevant questions as appropriate.

KEY POINTS ON RELATED AND COMPETING MEASURES

- NQF prefers endorsement of measures that assess performance for the broadest possible application (e.g., for as many possible individuals, entities, settings, and levels of analysis) for which the measure is appropriate, as indicated by the evidence
- The endorsement of multiple competing measures should be by exception, with adequate justification
- Harmonization of related measures should be done to the extent possible; differences in specifications should be justified

CHANGE LOG

DATE	CHANGES MADE
1/9/14	Updates to CSAC and Steering Committee processes
1/6/14	Updates to disclosure section, page 15- Ashlie- updated last updated date in header
1/3/2014	Additional information included in Staff Review paragraph on page 22.
12/30/2013	Section VI. Measure Evaluation Criteria add p.31-55.