# Committee Guidebook for the NQF Measure Endorsement Process

*VERSION 6.5*

*Last updated: August 2021*

## Contents

# The National Quality Forum

## Who is NQF?

The National Quality Forum (NQF), established in 1999, is a nonprofit, nonpartisan, membership-based organization. NQF 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 works 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 more than 400 organizational members that give generously of their time and expertise. In 2020, more than 650 individuals volunteered on more than 30 NQF-convened Committees, working groups, and partnerships. The NQF Board of Directors, which is composed of key public- and private-sector leaders who represent major stakeholders in America's healthcare system, governs the organization. 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. 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?

In 2002, 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. Over the years, NQF has assembled a portfolio of more than 550 NQF-endorsed measures—most of which are in use by both the private and public sectors—and an enormous body of knowledge about measure development, use, and performance improvement. NQF plays a key role in providing thought leadership to help shape our national health and healthcare improvement goals through the National Quality Partners. NQF also provides public input to the federal government and the private sector on optimal, aligned measure use via its convening of the Measure Applications Partnership (MAP). Additionally, NQF convenes multistakeholder groups to explore and make recommendations for diverse measurement science topics, thus informing the quality measurement enterprise.

NQF evaluates, 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 that physicians, other clinicians, and organizations can take to make healthcare safe and equitable;
- enhance transparency around quality and cost of healthcare;
- ensure accountability of healthcare providers; and
- generate data that help 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 healthcare 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. Both public and private payers use measures 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, important 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 find NQF-endorsed measures. This system allows users to search by measure title or number, as well as by condition, care setting, or measure steward, as well as by several other characteristics. QPS also allows users to provide feedback at any time about the use and usefulness of measures. QPS can also be used to learn from other measure users about how they select and implement measures in their performance improvement programs.

### Where do I find more information about NQF?

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

## 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 marketplace. 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, and improvement in symptoms or functioning. 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 healthcare 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 for patients with multiple chronic conditions** – Under the direction of the multistakeholder Multiple Chronic Conditions (MCCs) Standing Committee, NQF has developed a person-centric measurement framework for individuals with MCCs. Specifically, this framework defines 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.

**Measures that address healthcare disparities** – NQF's convening of the Disparities Standing Committee (DSC) complements our extensive work in quality measurement, providing a cross-cutting emphasis on healthcare disparities across all of NQF's work. Most importantly, the DSC developed a high-level roadmap for measuring and reducing disparities, which lays out how measurement and associated policies can be leveraged to promote health equity. In addition, the DSC provides expertise on emerging issues in measurement science and disparities (e.g., risk adjustment, stratification, and cross-cultural effects on patient surveys), provides advice and/or technical expertise on disparities to other committees (i.e., cross cutting committees or the MAP), and provides strategic direction and guidance to NQF and the measurement field on promoting health equity through measurement.

**Measures that are methodologically sound** – Healthcare performance measures are becoming increasingly more complex and sophisticated. NQF has recognized the need to collaborate with stakeholders who have deep methodological expertise to help standing committees evaluate measures for potential endorsement and to serve in an advisory capacity to NQF on methodologic issues related to measure testing, risk adjustment, and measurement approaches. To this end, NQF established the Scientific Methods Panel in 2017.

**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 not 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 a facilitative role to support measure

developers, NQF developed a process to ensure harmonization, address competing measures issues adequately, and provide enough time to resolve questions. The Harmonization Guidance and Definitions document provides an overview and outlines key definitions.

**eCQMs and health information technology (IT)** – 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 those data are easily shared between IT systems. NQF's health IT initiatives, which have 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. NQF has special processes and procedures for developers who are interested in submitting eCQMs (electronic clinical quality measures) for potential endorsement.

**Measures in nascent areas** – NQF continues to create needed strategic approaches, or frameworks, to measure quality in areas critical to improving health and healthcare for the nation but for which quality measures are too few, are underdeveloped, or nonexistent. A measurement framework is a conceptual model for organizing ideas that are important to measure for a topic area and for describing how measurement should take place. NQF's foundational frameworks identify and address measurement gaps in important healthcare areas, underpin future efforts to improve quality through metrics, and ensure safer, patient-centered, and cost-effective care that reflects current science and evidence. Recent efforts have resulted in frameworks for population-based trauma outcomes, healthcare system readiness, chief complaint-based quality for emergency care, feedback loops for performance measures, and patient safety in ambulatory settings.

The figure below is an illustrative example of the life cycle of a performance measure.

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.


## 4 RESULTS SPUR CHANGE IN PRACTICES

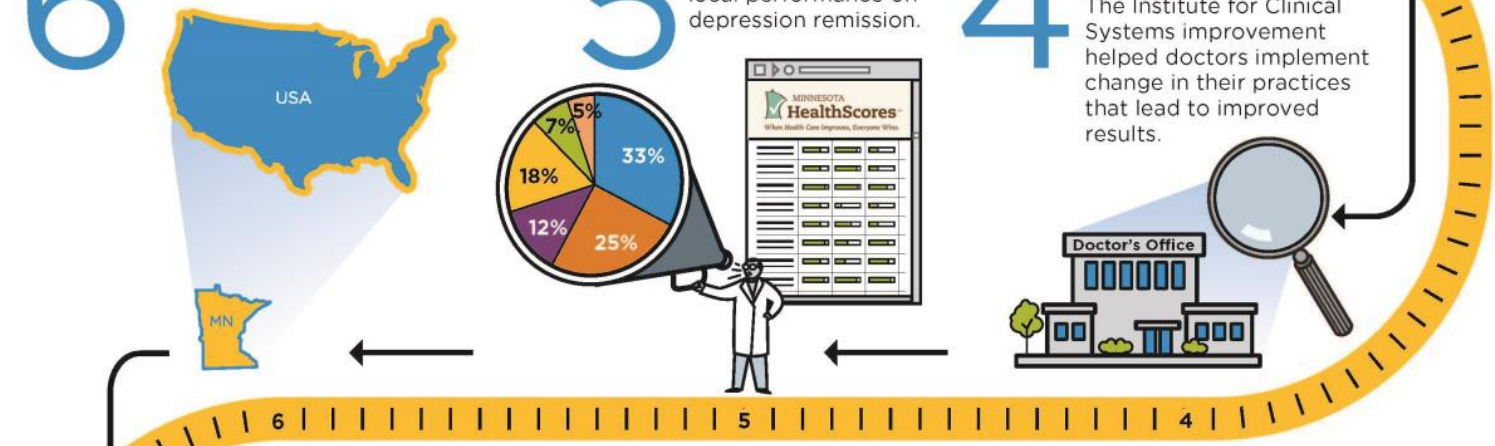
The Institute for Clinical Systems improvement helped doctors implement change in their practices that lead to improved results.

## 5 RESULTS MADE PUBLIC

MN Health Scores website publicly reports local performance on depression remission.

## 6 NATIONAL CONSENSUS STANDARD

NQF endorsed the measure as a national consensus standard.

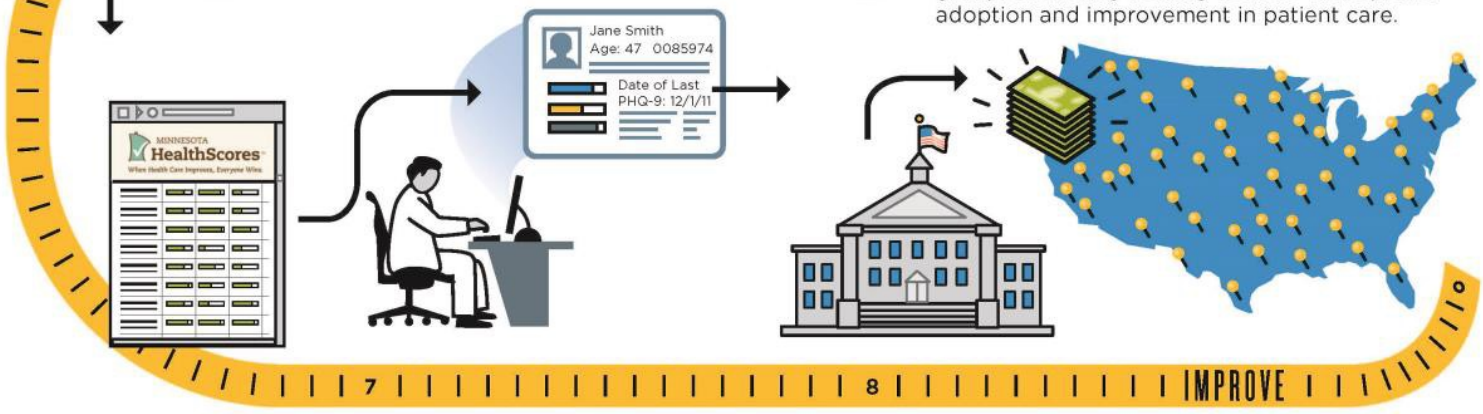



## 7 ELECTRONIC HEALTH RECORDS

MN Community Measurement retooled the measure for use in electronic health records.

## 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.




Source: National Quality Forum, MN Community Measurement 10-2012
Illustration: Funnel, Inc.

WWW.QUALITYFORUM.ORG

## 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 webpage 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

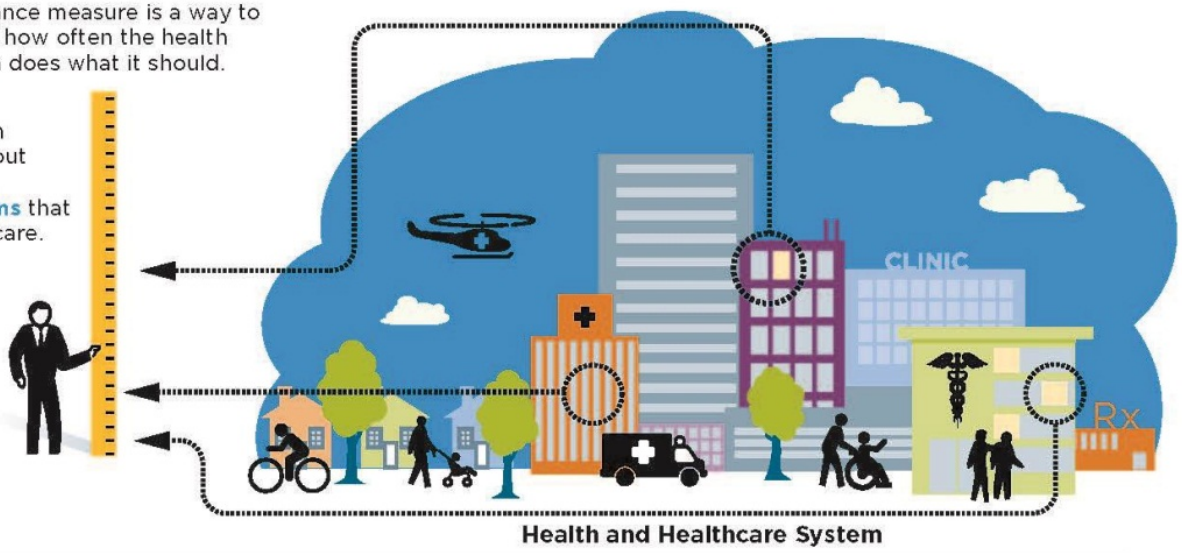
NATIONAL QUALITY FORUM

# 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.

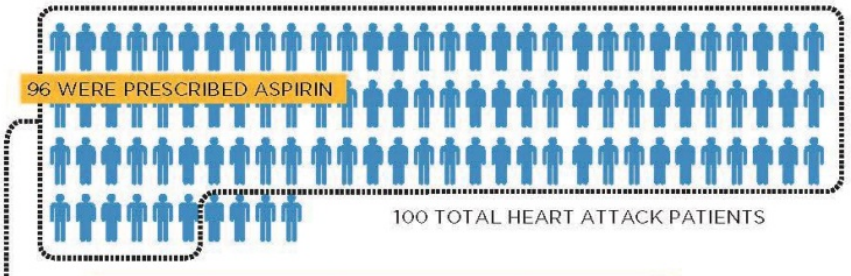
Measures are based on scientific evidence about **processes, outcomes, perceptions, or systems** that relate to high-quality care.




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.

# 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 20 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 six principal steps. Each contains several substeps and is associated with specific actions. The following section provides an overview of the six major steps of the CDP. Sections V and VI provide additional detail on how Standing Committees evaluate measures.

### *1. Intent to Submit*

Until recently, NQF issued a formal call for candidate standards prior to the start of the evaluation of standards in a particular topic area. Beginning in 2017, however, this has become a continuous call for measures across all topic areas, with pre-specified evaluation cycles and submission deadlines.

Each candidate measure has a measure steward who assumes responsibility for the submission of the measure for potential endorsement to NQF. Measure stewards or developers must notify NQF of their Intent to Submit at least three months prior to the desired designated cycle's measure submission deadline. This notification signals the measure steward's or developer's readiness for endorsement consideration and allows adequate opportunity for technical assistance prior to submitting measures for evaluation. As part of the Intent to Submit process, stewards or developers will submit **full measure specifications** to NQF and testing information to NQF, along with other information as needed (e.g., a feasibility assessment for eCQMs).

### *2. Call for Nominations*

Volunteer, multistakeholder Committees are central to the consensus process, and the success of NQF's projects is due in large part to the participation of its Standing Committee members. NQF currently convenes standing committees for various project topic areas. These Committees are responsible for evaluating both new measures and previously endorsed measures, as well as ad hoc and other project work in their topic areas. Other groups serve as an important adjunct to NQF's Standing Committees by helping to ensure broad representation on the standing committee and providing specific technical expertise when needed. Additional detail about these other expert groups is included at the end of the following section.

### *3. Candidate Consensus Standards (Measure) Review*

In addition to new measures, previously endorsed measures undergo evaluation for maintenance of endorsement approximately every three years. 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 maintenance of endorsement evaluation. To submit a measure for an initial endorsement evaluation or a maintenance-of-endorsement evaluation, a measure steward must complete and submit specific information about the measure via an online submission process through the NQF website.

The relevant Standing Committee conducts a detailed review of all submitted standards, sometimes with the help of a Technical Advisory Panel, expert reviewers, or members of other convened Standing Committees. During this review process, the Committee may meet several times, via webinars, 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 evaluation are described in Section VII of this guidebook.

### *4. Public Comment With Member Support*

Both NQF members and interested members of the public can submit comments on the measures during each evaluation cycle via the NQF website. One continuous public commenting period spanning at least 16 weeks allows NQF members and the public to provide comments on measures under evaluation at a time that is most convenient for them. This commenting period will open approximately 10 weeks prior to the Committee evaluation meeting and close 30 days after NQF posts the draft technical report on the NQF website. NQF will include all comments received at least four weeks prior to the committee evaluation meeting as part of the committee materials, for discussion by the Standing Committee during the evaluation meeting. After a Standing Committee completes its initial review of the submitted candidate standards, NQF posts a draft of the Committee's recommendations—or draft report—on the NQF website for review and comment by members of NQF and the public. This includes measures that were recommended for endorsement by the Standing Committee, those that were not recommended, and those for which consensus regarding endorsement was not reached. As part of NQF's commitment to transparency, all submitted comments will be posted on the NQF website, where any site visitor can review them. As of fall 2017, NQF members will have the opportunity to express their support ("Support" or "Do Not Support") for each measure to inform the Committee's recommendations during the commenting period. This expression of support (or not) during the commenting period replaces the member voting opportunity that was previously held subsequent to Committee deliberations.

All submitted comments are reviewed by the Standing Committee, and all submitted post-evaluation comments receive responses from the Standing Committee, measure developers, and/or NQF, as appropriate. The Standing Committee may revise its recommendations in response to a specific comment

or series of comments that are submitted during this phase of the CDP. It will revote on any measure for which consensus was not reached. All measures must have a recommended outcome after this meeting. A measure is recommended for endorsement by the Standing Committee when the vote margin on all must-pass criteria (Importance, Scientific Acceptability, Use), and overall, is greater than 60 percent of voting members in favor of endorsement. A measure that does not receive greater than 60 percent high/moderate ratings on all "must-pass" criteria AND the overall recommendation is not recommended for active endorsement.

### *5. Measure Endorsement Decision by the Consensus Standards Approval Committee (CSAC)*

The Consensus Standards Approval Committee (CSAC), an advisory Standing Committee appointed by the NQF Board of Directors, is the governing body that has the most direct responsibility for overseeing the implementation of NQF's CDP. The work of the CSAC focuses on NQF's evaluation criteria, the provision of endorsement decisions on 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 expression of support then decide whether to do the following:

- Endorse a measure that a Standing Committee recommends for endorsement
- Uphold the Standing Committee's recommendation to not endorse the measure
- Disagree with the recommendation of the Standing Committee and return the measure back to the Committee for reconsideration

### *6. Measure Appeals*

Following the CSAC's endorsement decisions, measures will enter a 30-day Appeals period. Any party may request an appeal of a CSAC decision to endorse or not endorse a measure, except in the case where a Standing Committee does not recommend a measure for endorsement and the CSAC concurs. CSAC decisions to endorse a measure with reserve status or approve a measure for trial use are not appealable. For an appeal to be considered, it must include information that clearly demonstrates there was a procedural error that is reasonably likely to affect the outcome of the original endorsement decision, or there is new information or evidence that was unavailable at the time the CSAC made its endorsement decision that is reasonably likely to affect the outcome of that decision. Appeals can be requested through appeals@qualityforum.org or the NQF Measure Database. All appeals are published on the Appeals Board section of the NQF website. Appeals will be made to an Appeals Board, composed of NQF Board members and former CSAC and/or Committee members. The Appeals Board will adjudicate appeals to measure endorsement decisions without a review by the CSAC. The decision of the Appeals Board will be final.

## Standing Committee Composition, Roles, and Responsibilities

### Standing Committee Application Requirements

NQF invites nominations for standing committees on an annual basis. Staff will publicize details regarding the desired perspectives or expertise for new Committee members at that time. Self-nominations are

welcome. Third-party nominations must indicate that the individual has been contacted and is willing to serve. All nominations remain active for one year. To be considered for appointment to a Standing Committee, individuals must provide 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 a nomination 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.

## Composition of Standing Committees

Topical Standing Committees include 20 to 25 individuals representing 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. For larger topic areas that include multiple conditions or cross-cutting areas, NQF will utilize technical expertise and/or expert reviewers for specific areas as needed (described more at the end of this section).

Nominations are for 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 and solicit input from colleagues and fellow stakeholders throughout the process.

## Standing Committee Terms

New Standing Committee members are appointed to a three-year term with the ability to extend for one additional term of two years. After two consecutive terms, Committee members must step down for a full term (three years) before becoming eligible for reappointment. NQF reserves the right to make an exception to the policy above if one-third or more of Standing Committee members are scheduled to roll off at the same time. A Standing Committee member's term begins on January 1st after selection to the Standing Committee, following the close of the roster commenting period. Standing Committee co-chairs are appointed to a three-year term with an option to extend for two additional two-year terms.

## Standing Committee Expectations and Time Commitment

Participation on a Standing Committee requires a significant time commitment. Committee members are expected to participate in all currently scheduled meetings. Over the course of the term, additional

meetings will be scheduled, or meetings may be rescheduled; new dates are set based on the availability of the majority of the Committee.

Committee participation includes the following (estimated times may vary depending on the number and complexity of the measures under review, as well as the complexity of the topic and multistakeholder consensus process):

- Participate in the scheduled orientation call (two hours)
- Identify and disclose potential biases (real or perceived)
- Review all measure submission forms (approximately two hours per measure)
- Complete all surveys and evaluations
- Attend all scheduled evaluation meetings. These may be in-person meetings (one to two full days in Washington, D.C.) or a series of webinars (typically two hours each)
- Lead discussion of some measures at calls or meetings and participate in the discussion and vote on ratings and recommendations for all measures
- Review meeting summaries and/or draft reports
- Complete measure evaluation by reviewing the comments received on the draft report and then participate on the post-comment webinar (two hours)
- Complete additional measure evaluations by conference call or webinar if needed
- Participate in additional calls or webinars as necessary
- Present measures and lead discussions for the committee on conference calls, webinars, and other meetings

*Committee members are responsible for notifying the NQF project team if he/she:*

- Changes employers or contact information
- Is unable to attend a scheduled meeting
- Has a prolonged conflict that emerges during his/her term that will interfere with meeting the obligations of standing committee membership, in order to determine whether ongoing membership on the Committee is warranted or if "inactive" status can be granted for a cycle

*If a member has poor attendance or participation:*

- The NQF staff will contact the member and ask if he/she would like to resign
- NQF Staff reserves the right to remove any member for persistent poor attendance or lack of participation.

*If a member is unable to fulfill his/her term (for any reason):*

- NQF will identify a replacement through the pool of expert reviewers. If a replacement cannot be identified from the expert reviewer pool, NQF staff will review the nominations received during the most recent call for nominations.

- NQF staff will contact the potential replacement.

- Upon acceptance of Committee appointment, the new member would complete the term of the individual who was replaced.
- The outgoing member may not select a substitute to carry out the remainder of the term.

## Standing Committee Disclosures of Interest

Per the NQF Conflict of Interest Policy for CDP Standing Committees, all nominees 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. This general DOI form must be completed annually in order to participate in measure evaluation.

Once nominees have been selected to serve on a Committee, a *measure-specific* DOI form will be distributed near the beginning of each evaluation cycle. This measure-specific DOI is used to determine whether any members 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. Because Standing Committee members are asked to review various types of measures throughout their term of service, NQF asks members to complete the measure-specific DOI for all measures being evaluated in each cycle , as well as any measures that are related to, or competing with, measures being evaluated, to ensure any potential conflicts or biases have been identified. Committee members who fail to return a completed measure-specific DOI prior to measure evaluation meetings will not be allowed to participate in the discussion or submit votes on the measures being evaluated.

## Role of the Standing Committee

The Standing Committee acts as a proxy for the NQF multistakeholder membership. The individual members 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" and have different points of view, members should use their own personal experience and expertise while serving on the Committee.

The primary responsibility of the Committee is to evaluate candidate measures using NQF's standard measure evaluation criteria. It also will consider national and CMS priorities 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 to get to know fellow committee members.

## Role of the Committee Co-Chairs

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

- Co-facilitate meetings, along with project staff (all co-chairs should participate in facilitation training hosted by NQF)

- 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
- Represent the Committee at the CSAC meetings or calls

## Additional Expertise

As noted earlier, other groups of experts serve as an important adjunct to NQF's Standing Committees as needed. These groups include a pool of expert reviewers, technical advisory panels, and the Scientific Methods Panel.

### *Expert Reviewers*

NQF's pool of expert reviewers serve as an important adjunct to NQF's Standing Committees. They are "on call" for the explicit use of CDP Standing Committees. An expert reviewer cannot serve on any other type of committee without entering the nomination process. All expert reviewers will adhere to the Standing Committee Policy (e.g., terms, conflict of interest, etc.) and are required to disclose any conflicts of interests, similar to the requirements of the members of the Standing Committee. Expert reviewers will remain in the pool until their term expires. Expert reviewers will only be required to fill out a measure-specific DOI if they are seated on the standing committee for a particular measure review cycle.

NQF anticipates the role of the expert reviewer to evolve over time, but the current role is as follows:

***EVALUATE MEASURES***

Expert reviewers will provide expertise (as needed) to evaluate the measures submitted for endorsement consideration. In order to evaluate measures, the expert reviewer will become a member of the seated Standing Committee for a specific amount of time (e.g., one cycle or longer, depending on the reason the expert reviewer was moved to the Standing Committee), such as:

- replacing a Committee member who becomes inactive;
- replacing a Committee member whose term has ended; and
- providing expertise that is not currently represented on the Committee to review submitted measures

The expert reviewers will vote on the measures under consideration. Additionally, the expert reviewers will be added to the Standing Committee roster, posted on the NQF topic area webpage for the particular cycle in which the reviewer is on the Committee.

***PROVIDE COMMENT/FEEDBACK ON MEASURES***

Expert reviewers who are not a part of the Standing Committee may provide comments and feedback on the measures throughout the measure evaluation process for the topic area to which they have been assigned. Expert reviewers who are not a part of the Standing Committee will not have the ability to vote on the measures submitted for endorsement consideration. NQF staff will alert expert reviewers to upcoming Committee activities. Expert reviewers are encouraged (but not required) to attend the meetings and provide commentary during public comment.

### *ENGAGE IN STRATEGIC DISCUSSIONS*

In the event no measures are submitted for endorsement consideration in a particular cycle, the Standing Committee will have an opportunity to discuss overarching measurement issues specific to their topic area. NQF staff will alert expert reviewers to these meetings. Expert reviewers are encouraged (but not required) to attend and participate in the meetings.

#### TECHNICAL ADVISORY PANELS

Based upon the expertise present on a Standing Committee, and through the pool of expert reviewers, a technical advisory panel may also be seated to provide needed expertise for the endorsement process. Members of a technical advisory panel are experts in their field. They provide guidance to a Standing Committee around specific technical issues related to some or all of the consensus standards being evaluated. At the direction of the Standing Committee, members of a Technical Advisory Panel may be charged with reviewing the evidence supporting candidate consensus standards and/or completing other reviews requiring technical expertise. Members of a technical advisory panel are selected primarily for their content expertise and experience. Members are also selected based upon their potential contribution to the project and the need for input from a particular stakeholder perspective. NQF members and the public will be given an opportunity to comment on Technical Advisory Panel rosters.

#### SCIENTIFIC METHODS PANEL

The Scientific Methods Panel was initiated in 2017 to conduct evaluation of new complex measures for the criterion of *Scientific Acceptability* (i.e., the reliability and validity subcriteria) and to serve in an advisory role to NQF on measurement science topics. Members of the Panel are expert methodologists. They are appointed to an initial two- or three-year term, with an optional two-year term to follow. In 2019, NQF expanded the size of the panel by adding an additional eight members. Going forward, NQF will issue an annual Call for Nominations for the Scientific Methods Panel in order to fill vacated seats. Additional information about Panel, including the roster, charge, and FAQs, is available on the Scientific Methods Panel webpage on NQF's public website.

NATIONAL QUALITY FORUM

# Multi-stakeholder Review: Criteria for Evaluating a Measure

## MULTI-STAKEHOLDER COMMITTEES OVERSEE ENDORSEMENT

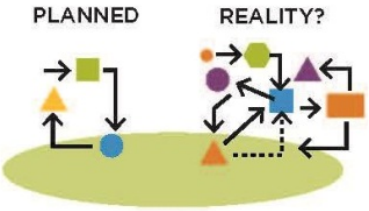
These committees evaluate measures by clinical condition against agreed upon criteria. Measures reviewed are endorsed and receive the NQF seal of approval. In order to receive NQF endorsement, measures must meet all five endorsement criteria.


### 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.

# The Measure Evaluation Process

## Before the Evaluation Meeting

### *Measures Submitted for Evaluation*

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 for endorsement are met. The questions included as part of the measure submission form are posted on NQF's website.

**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.
- If you have difficulty accessing the SharePoint site, please contact the NQF project staff.

### *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 standard 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 efficient, high quality healthcare for patients.

To determine whether a candidate measure should be endorsed by NQF, the Standing Committee evaluates the candidate measure 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 various entities 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 conduct a preliminary analysis of the measure information and assign a preliminary rating for the major evaluation criteria/subcriteria, in order to assist the Committee in its evaluation. For complex measures, NQF's Scientific Methods Panel will provide an analysis and rating of reliability and validity. Preliminary analyses will summarize key points of the submission as they pertain to NQF's evaluation criteria, provide hyperlinks so that readers can easily locate particular sections of the submission, and, when appropriate, provide additional context or interpretation for certain aspects of the submission (e.g., verifying that a testing methodology is appropriate). The preliminary ratings, which are included as part of the preliminary analysis information, are not binding, but instead, are meant to serve as input for Committee discussion.

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

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

- **Importance to Measure and Report** (this is not the same as "important to do") – Extent to which the specific measure focus is evidence based and important to making significant gains in healthcare quality 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.

- **Scientific Acceptability of the Measure Properties: Reliability and Validity** – 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, and 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. As of fall 2017, the "use" requirements under Usability and Use are must-pass for measures undergoing endorsement maintenance.

- **Comparison to Related and Competing Measures** – If a measure meets the above criteria and there are previously-endorsed or new related measures (i.e., have either the same measure focus or the same target population) or competing measures (i.e., have 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 for measure development, increased data reporting requirements, and confusion when measures produce conflicting results.

For each of the standard criteria, several subcriteria 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 are discussed in more detail in Section VII.

Committee members first review and evaluate the measures individually, 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 thus may not feel qualified to evaluate all aspects of a measure. All Committee members should contribute to the evaluation to the best of their abilities, knowing that the final evaluation rating and recommendation will be made by the full Standing Committee.

### *Scientific Methods Panel Evaluation*

The Scientific Methods Panel (SMP) was established in 2017 in response to recommendations that emerged from NQF's 2017 Kaizen process improvement activity. The SMP is a Standing Committee composed of statisticians, economists, methodologists, psychometricians, and others with like expertise. Its charge is to evaluate complex measures submitted to NQF against the Scientific Acceptability Criteria (i.e., reliability and validity). The SMP also is charged with providing guidance to NQF on the curation of the scientific acceptability criteria in an evolving quality measurement landscape, as well as providing guidance to the field on how best to demonstrate and represent that measures have met these criteria.

By transferring the evaluation of the scientific acceptability for some measures from Standing Committees to the SMP, NQF envisioned that patients, consumers, and caregivers on Committees can participate more fully in the discussions, that consistency in the evaluation of complex measures would improve, and the workload of the Standing Committees would be reduced.

The SMP evaluates reliability and validity of new complex measures, as well as complex maintenance measures with updated testing. Complex maintenance measures without updated testing can also be reviewed by the SMP at the discretion of NQF project staff. NQF defines complex measures as:

- Outcome measures, including intermediate clinical outcomes
- Instrument-based measures (e.g., PRO-PMs)
- Cost/resource use measures
- Efficiency measures (those combining concepts of resource use and quality)
- Composite measures

The SMP applies the NQF criteria for reliability and validity and provides ratings for each. NQF staff compile the ratings and then provide as summary of the SMP's evaluations to the relevant Standing Committee.

#### Consideration of Measures That Pass the SMP Evaluation or Where Consensus Is Not Reached:

Measures that receive a majority passing vote by the SMP, or measures for which the SMP has not reached consensus, will be forwarded to the Standing Committee for review and discussion at the relevant evaluation meeting. The staff preliminary analysis, which is provided to Standing Committee members, will include a summary of the SMP's ratings and evaluation. Standing Committees can determine whether to accept the SMP's ratings, or they may choose to re-adjudicate the criteria and submit its own vote.

Standing committees may be particularly interested in offering their own votes for the validity criteria, as clinical perspectives may further inform the evaluation of the criterion.

### Consideration of Complex Measures That Did Not Pass the Scientific Methods Panel's Evaluation

Measures that receive a majority non-passing vote by the SMP for **either** reliability or validity are considered to have not passed the SMP's evaluation. Beginning with the fall 2019 evaluation cycle, Standing Committee members will have an opportunity to pull measures such measures for further discussion. In some cases, Standing Committees may also have the option to revote on the criteria. Details of the SMP evaluation and will be shared with the Standing Committee in advance of the evaluation meeting in order to give Committee members an opportunity to identify which measures should be added to the agenda for discussion by the full Standing Committee. Any measure that is not pulled by a Standing Committee member does not advance to the Committee for discussion. Endorsement status **will be removed automatically from maintenance measures that do not pass the SMP, if they are not pulled for discussion**. Any measure that is pulled will be discussed during the evaluation meeting. The Committee member who pulls the measure for discussion will be assigned as lead discussant for the measure.

Some measures that do not pass the SMP evaluation may be eligible standing committee re-vote on the scientific acceptability criteria. Measures are eligible for potential revote if the measure did not fail for one of the following reasons:

- Inappropriate methodology or testing approach applied to demonstrate reliability or validity
- Incorrect calculations or formulas used for testing
- Description of testing approach, results, or data is insufficient for SMP to apply the criteria
- Appropriate levels of testing not provided or otherwise did not meet NQF's minimum evaluation requirements

### *NQF Staff Preliminary Analysis and Measure Worksheet*

NQF staff will review the submitted information against the NQF criteria as part of a "preliminary analysis" of the measure. As part of the preliminary analysis, staff will highlight areas for discussion by the committee, as well as any questions or critical decisions for the committee to consider and offer a preliminary rating for the major evaluation criteria/subcriteria. For measures evaluated by the SMP, NQF staff will provide a document that combines the initial evaluations of the SMP members assigned to that measure. Staff also will provide a summary of the SMP's evaluation; this will be included with the staff preliminary analysis. The preliminary analysis, along with materials submitted by the measure developer/steward, will be packaged into a "measure worksheet," which the Standing Committee will use to evaluate each measure. The measure worksheet also will include any feedback from the field on implementation or use of the measure, and any measure-specific comments submitted to NQF prior to the Committee's evaluation. Internal links will be embedded in the measure worksheets to allow for easy navigation throughout the document.

### *Initial Evaluation by Individual Committee Members*

When conducting the initial in-depth evaluations, each Committee member will consider all measures in light of all criteria and subcriteria. A survey tool will be used to collect Committee members' initial thoughts about the measures for further discussion. Committee members should **contact staff for**

**assistance if encountering difficulty using the survey tool.** Initial evaluation responses will be made available to all standing committee members and measure developers prior to the evaluation meeting.

To facilitate the Committee discussions, two to three discussants will be designated for each measure. These 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 full Committee evaluation;
- begin the discussion of the measure evaluation including
  - presenting a brief description of the measure
  - summarizing the evaluation of each criterion based on all of the evaluation comments entered through the SharePoint survey tool, highlighting areas of concern or difference of opinion and the issues or questions posed in the preliminary analysis
  - verbalizing conclusions regarding how well the measure meets NQF's evaluation criteria (refer to Section VII on the criteria).

### *Evaluation by the Entire Committee at the In-Person Meeting or Web Meeting*

NQF Standing Committees will meet either in person or via web meeting(s) to evaluate measures and make recommendations. For meetings that are conducted in-person, committee members are requested to bring laptop computers or tablets capable of viewing the documents electronically. An internet connection will be available.

- The following reflects context and activities that occur during evaluation meetings (whether done via webinar or in person). Transparency – In-person/web meetings are open to the public (in person and by phone). The proceedings are recorded a meeting summary is posted on the project website.

- Disclosure of interests – During introductions at the beginning of the meeting, each Committee member is asked to disclose any interests as identified on the disclosure of interest form.

- Measure developers are encouraged to be present during the meeting (in person or via phone) to respond to any issues or questions. Measure developers are given an opportunity to introduce 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 discussants will introduce each measure and begin the discussion. After discussion by the entire Committee, a vote is taken on each criterion and on selected subcriteria, and lastly, 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 when the vote margin on all must-pass criteria (Importance, Scientific Acceptability, Use), and overall, is greater than 60 percent of voting members in favor of endorsement. A measure is not recommended for endorsement when the vote margin on any must-pass criterion or overall is less than 40 percent of voting members in favor of endorsement. An exception is if a measure passes every criterion but performance gap. In this case, the Standing Committee may choose to recommend the measure for inactive endorsement with reserve status. The Standing Committee has **not** reached consensus if the vote margin on any must-pass criterion or overall is between 40 and 60 percent, inclusive, in favor of endorsement.

    - When the Standing Committee has not reached consensus, all measures for which consensus was not reached will be released for NQF member and public comment. It will consider the comments and revote on those measures during a webinar convened after the commenting period closes.

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

### *Committee Ground Rules for Meetings*

Committee members act as a proxy for NQF's membership. As such, this multistakeholder 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.

Evaluation meeting agendas typically are 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 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 Evaluation Meeting

After a project's Standing Committee completes its initial evaluation of the submitted measures (including, if needed, a post-meeting call convened after the main evaluation meeting), a draft of the recommendations—or "draft technical report"—is posted on the NQF website for review and comment by 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, on the event calendar, and on the specific project page. NQF also sends 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 recommendations 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. NQF members also will have the opportunity to express their support ("Support" or "Do Not Support") for each measure to inform the Committee's recommendations. All NQF member organizations are eligible to express support/non-support of all measures being considered for endorsement or re-endorsement.

Comments received during the comment period provide feedback to the Standing Committee on the measures themselves, as well as on the Committee's evaluation and recommendations for endorsement. NQF members and nonmembers 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 member expressions of support/non-support and adjust any recommendations as needed.**

### *Developer Request for Reconsideration by the Standing Committee of a Measure Not Recommended*

Requests for reconsideration related to appropriate application of the criteria are submitted through the public and member comment process. The requestor must cite the specific evaluation criteria or subcriteria thought to be applied improperly to the specific information as submitted to, and evaluated by, the Standing Committee. It will review the cited information in the submission form, any additional information, and the criteria under question during the comment review process, with the option to revote on the measure.

### *Post-Comment Conference Call*

After the conclusion of the member and public comment period, the Standing Committee will meet via conference call to review all comments not already considered during the evaluation meeting. The Standing Committee may seek 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, the revised version of the draft report may be recirculated for a second comment period for NQF members and the public. If a revised version of the draft report is recirculated for a second comment period, the review will follow the same process as the initial review and comment period.

### *Consensus Standards Approval Committee (CSAC)*

The Consensus Standards Approval Committee (CSAC), an advisory Standing Committee of the NQF Board of Directors, is the governing body that has the most direct responsibility for overseeing the implementation of NQF's CDP. The work of the CSAC focuses on NQF's evaluation criteria, the provision of endorsement decisions for 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. The CSAC membership includes a simple majority of consumers and purchasers. Some CSAC members possess specific expertise in measure development, application, and reporting.

The CSAC holds two in-person meetings annually and, otherwise, convenes monthly by conference call. During selected meetings, the CSAC reviews the recommendations of the Standing Committee, the public and member comments and the responses, and the results of NQF member expression of support, prior to making its decision regarding endorsement. All CSAC meetings wherein measures are discussed are open to NQF members and the public, and audience members have the opportunity to comment on the measures under consideration. Standing Committee co-chairs attend the call to represent the Committee at the CSAC meetings. Measure developers are expected to attend the relevant CSAC meetings in which their measures are being considered for endorsement and to answer any questions from members of the CSAC. Information about each CSAC meeting is available on the NQF website, including the meeting's agenda and materials and the physical location and dial-in information.

Following the opportunity for public comment, the CSAC will make final measure endorsement decisions (ratification by NQF's Board of Directors is no longer required for endorsement). The CSAC will review Standing Committee recommendations and then decide whether to do the following:

- Endorse a measure that a standing committee recommends for endorsement
- Uphold the Standing Committee's recommendation to not endorse the measure
- Disagree with the recommendation and return the measure back to the committee for reconsideration

#### CSAC Criteria for Decision Making

To ensure a consistent approach to endorsement decisions, the CSAC identified the following overarching guidance and criteria to guide its decision making:

##### OVERARCHING GUIDANCE

To ensure a consistent approach to endorsement decisions, the Consensus Standards Approval Committee (CSAC) identified the following criteria to guide its decision making. As a general principle, the CSAC should not re-adjudicate or overturn a Standing Committee's endorsement recommendation, but rather

determine if there is consistency in the rationale used by Standing Committees when recommending measures. The CSAC, however, may send a measure back to a Standing Committee for reconsideration if there are concerns with any of the rationale/criteria below. These concerns will be documented and communicated to the Standing Committee and the public.

#### DECISION MAKING CRITERIA

- **Strategic importance of the measure.** The CSAC will consider the value-add 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. To assess additive value and importance, the CSAC should consider NQF's measure selection attributes including outcome-focused, high opportunity for improvement, patient and caregiver focus, support integrated view of care, reasonable data collection burden, and impact/prevalent condition.
- **Cross-cutting issues concerning measure properties.** The CSAC will consider whether criteria concerning measure properties are consistently and appropriately applied across the entire portfolio.
- **Consensus Development Process concerns**. The CSAC will consider all concerns raised during the CDP by all stakeholders, such as sufficient attention to member and public comment. CSAC may conclude that additional efforts should be made to address these concerns before making an endorsement decision on the measure (e.g., returning a measure to the Standing Committee for reconsideration).

### CSAC Voting

The meeting quorum for all CSAC meetings is 100 percent of the CSAC body. Greater than 60 percent approval for endorsement of a measure by voting CSAC members is required to grant endorsement. The CSAC does not have a consensus-not-reached threshold. The CSAC is expected to vote on candidate standards in two ways. If a project is presented at an in-person meeting, the CSAC is expected to vote during the in-person meeting. If a project is presented during a conference call, the CSAC is expected to vote online; in this case, the CSAC has seven calendar days to submit votes.

### Developer Requests for Reconsideration by the CSAC of a Measure Not Recommended

All reconsideration requests related to the criteria must go to the Standing Committee, as described above. However, a measure developer also may request reconsideration by the CSAC co-chairs if the developer believes that NQF's CDP was not followed. Developers must send a written request for reconsideration to the CSAC co-chairs at least two weeks prior to the CSAC call/meeting that grants endorsement, citing the issues within a specific CDP process step, how it was not followed properly, and how it resulted in the specific measure not being recommended.

#### PROCESS FOR CSAC REVIEW

- Staff will prepare a summary of the CDP process for the measure(s), with special attention to the issues raised and Committee's discussion and explanation, in the reconsideration request.
- The CSAC co-chairs may:
  - uphold the Standing Committee final recommendation if the process was followed;
  - ask for input from the CSAC, particularly if co-chairs think there is merit to the assertion of not following the CDP;

- request additional expert input; or
- if a breach in the CDP was identified, determine if it may have adversely affected the outcome for the specific measure.

- If the CSAC co-chairs determine that a breach in the CDP occurred that may have adversely affected the outcome of the specific measure, then the entire CSAC will evaluate the circumstances and determine a course of action on a case-by-case basis.

### *Appeals*

Following the CSAC's endorsement decisions, measures will enter a 30-day Appeals period. Any party may request an appeal of a CSAC decision to endorse or not endorse a measure, except in the case where a Standing Committee does not recommend a measure for endorsement and the CSAC concurs. CSAC decisions to endorse a measure with reserve status or approve a measure for trial use are not appealable. For an appeal to be considered, it must include information that clearly demonstrates there was a procedural error that is reasonably likely to affect the outcome of the original endorsement decision, or there is new information or evidence that was unavailable at the time the CSAC made its endorsement decision that is reasonably likely to affect the outcome of that decision. Appeals can be requested through appeals@qualityforum.org or the NQF Measure Database. All appeals are published on the Appeals Board section of the NQF website.

The Appeals Board, composed of NQF Board members and former CSAC and/or committee members, will adjudicate appeals to measure endorsement decisions, without a review by the CSAC. The decision of the Appeals Board will be final.

Throughout the process, project staff will serve as liaisons between the CSAC, the Appeals Board, the Committee, developers/stewards, and the appellant(s) to ensure the communication, cooperation, and appropriate coordination to complete the project efficiently.

## Measure Evaluation Criteria

For details on NQF's measure evaluation criteria and guidance on how to evaluate measures based on these criteria, please refer to the NQF's Measure Evaluation Criteria and Guidance for Evaluating Measures for Endorsement.

### 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 and performance gap subcriteria help to answer the question about whether and how a measure is important to measure and report. Most of the standard 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)**
   a. Evidence to Support the Measure Focus (must-pass)
   b. Performance Gap, including Disparities (must-pass)
   c. For composite measures: Quality Construct and Rationale (must-pass)
2. **Scientific Acceptability of Measure Properties (must-pass)**
   a. Reliability [includes additional subcriteria] (must-pass)
   b. Validity [includes additional subcriteria] (must-pass)
   c. For composite measures: Empirical Analysis Supporting Composite Construction (must-pass)
3. **Feasibility**
   a. Required data elements routinely generated and used during care delivery
   b. Availability in electronic health records or other electronic sources OR a credible, near-term path to electronic collection is specified
   c. Data collection strategy can be implemented
4. **Usability and Use**
   a. Use (must-pass for maintenance measures)
      i. Accountability and transparency
      ii. Feedback on the measure by those being measured and others
   b. Usability (not must-pass for maintenance measures)
   c. Improvement
   d. The benefits to patients outweigh evidence of unintended negative consequences to patients
5. **Comparison to Related or Competing Measures**
   a. Measure specifications are harmonized **OR** differences are justified
   b. 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, 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 and reflect variation in performance, overall, less-than-optimal performance, or disparities. 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 this criterion and both its 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 may 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, as well as 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 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 not be endorsed. In fact, Feasibility may improve with broader implementation, and ways to improve Feasibility should be sought for important and scientifically sound performance measures.

The fourth criterion is 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 use—and then implemented—in various programs, the Usability and Use criterion is not designated as must-pass for initial endorsement. However, the subcriteria under Use (4a1 and 4a2) will be must-pass when evaluating measures for continued endorsement.

Lastly, if the Standing Committee agrees that a measure has met the first four NQF evaluation criteria and is suitable for endorsement, 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 is not a "best" measure, endorsing measures that are consistent to the extent possible.

### *Rating Scales*

Usually, the evaluation of a measure is not a straightforward yes/no or all-or-nothing determination. Instead, measures typically meet the criteria to a greater or lesser extent. As a result, 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 that the information submitted 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 or information.

### *Evaluating New Versus Previously Endorsed Measures*

All measures—both new and previously endorsed measures that are undergoing maintenance of endorsement—are expected to meet current evaluation 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 maintenance of endorsement evaluation, NQF expects measures to be in use—and therefore developers should submit data from implementation of the measure as specified to demonstrate performance gaps (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 subcriterion.

## Closer Look at NQF's Evaluation Criteria and Subcriteria

This section provides 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 titled What Good Looks Like. Additional detail regarding NQF's evaluation criteria and guidance can be found in various reports available on the Submitting Standards webpage (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 and important for making significant gains in healthcare quality 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 many things 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 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 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 healthcare performance 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 by empirical evidence to desired outcomes.

**Key Points for Evaluating Importance to Measure and Report**

- 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 (including patient-reported process or structural measures) that are closest to desired outcomes.
- Both subcriteria under Importance to Measure and Report are "must-pass"; therefore, each should 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 many 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 outcome, 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.

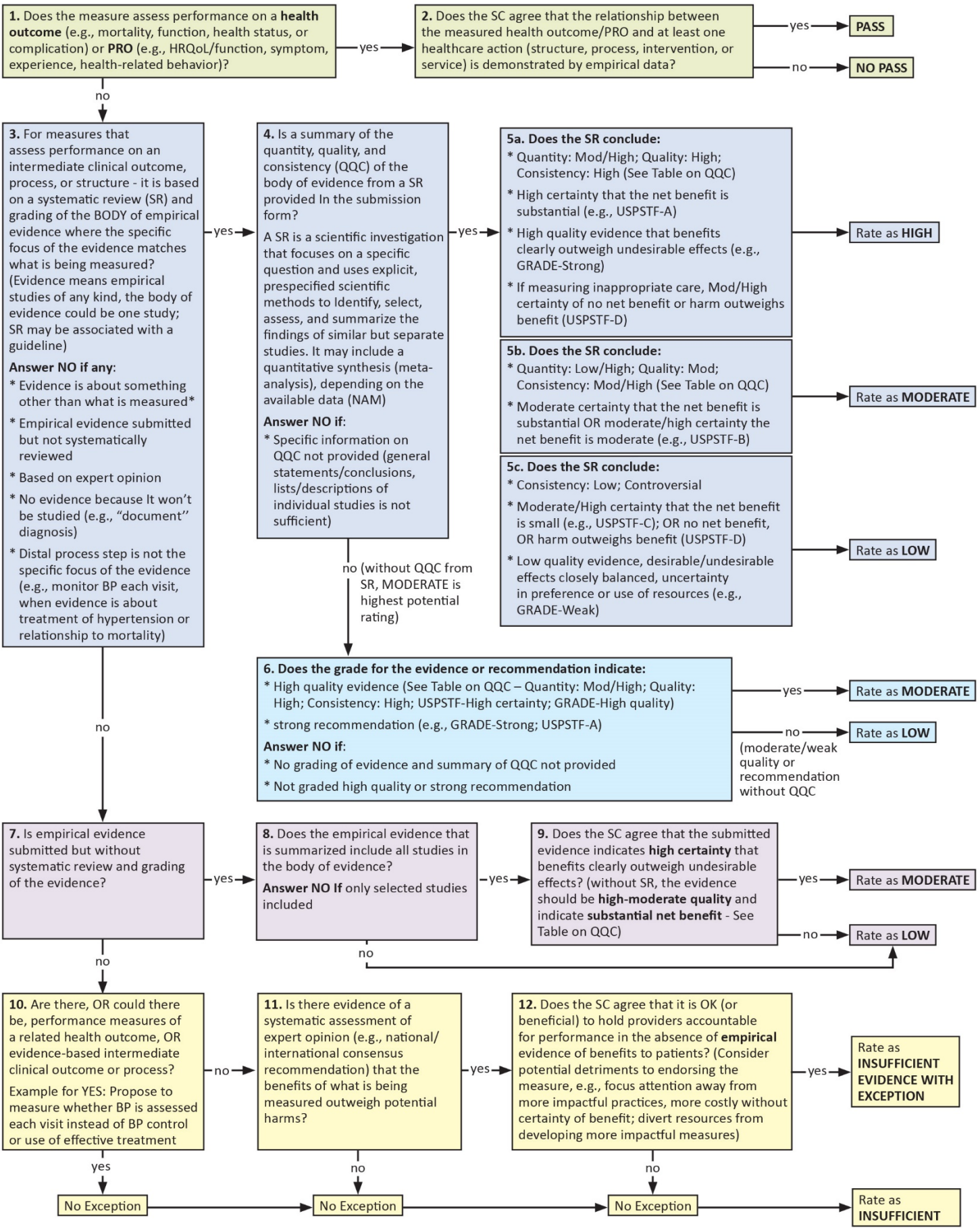
Evidence refers to empirical studies but is not limited to randomized controlled trials. The preferred sources of evidence are systematic reviews and grading of a body of evidence that are conducted by independent organizations (e.g., USPSTF, Cochrane Collaboration, etc.). Because not all healthcare 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 provide empirical data to demonstrate the relationship between the outcome and at least one healthcare structure, process, intervention, or service. If such empirical data are not available, a demonstration of wide variation in performance can be used as evidence for the outcome measure instead, as such data implies differences in practices that result in differences in performance, assuming the data reflect a robust number of providers with results that are not subject to systematic bias.

A final evidence requirement—limited to measures derived from patient report—is that the target population (e.g., dying patients, hospitalized patients, etc.) values the measured outcome, process, or structure and finds it meaningful. This could be demonstrated, for example, by describing how patient input was included in the development of the underlying instrument used to collect the patient-reported data or if focus groups of patients agree on the value of the performance measure.

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

## Algorithm 1. Guidance for Evaluating the Clinical Evidence




**NOTE:** Submissions for instrument-based measures that are based on patient report must include information to demonstate that the target population values the measured structure, process, or outcome and finds it meaningful. If not demonstrated, then rating should be INSUFFICIENT.

**Key Points for 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 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.

**Table 1. Evaluation of Quantity, Quality, and Consistency of the 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 |

| **Definition/ Rating** | **Quantity** of Body of Evidence | **Quality** of Body of Evidence | **Consistency** of Results of 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 |
| Moderate | 2-4 studies[b] | • Non-RCTs with control for confounders that could account for other plausible explanations, with large, precise estimate of effect<br>OR<br>• 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<br><br>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<br>OR<br>• 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<br>OR<br>• wide confidence intervals prevent estimating net benefit<br>If only one study, then estimated benefits do not greatly outweigh harms to patients |

| Definition/ Rating | **Quantity** of Body of Evidence | **Quality** of Body of Evidence | **Consistency** of Results of Body of Evidence |
|---|---|---|---|
| Insufficient to Evaluate | • No empirical evidence<br>OR<br>• Only selected studies from a larger body of evidence | • No empirical evidence<br>OR<br>• Only selected studies from a larger body of evidence | No assessment of magnitude and direction of benefits and harms to patients |

○ aStudy 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.

○ bThe 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 opportunity for improvement). Opportunity for improvement can be demonstrated by 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 previously would not be considered as meeting the performance gap subcriterion and thus might not be granted continued endorsement. However, the Standing Committee may, in certain circumstances, recommend those measures for Inactive Endorsement With Reserve Status ("Reserve Status"). Use of the Reserve Status should 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) (see Reserve Status policy).

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

**Table 2. Generic Scale for Rating Subcriteria 1b, 1c, 2c, and Criteria 3 and 4b**

| 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, not relevant, responsive, or specific to the particular question) |

**Key Points for Evaluating 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; and
  - data on disparities

### Subcriterion 1c: 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 in the section on composites measures in the Measure Evaluation Criteria and Guidance document. 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)

- relationships among the component measures (e.g., whether they are correlated or not, processes that are expected to lead to better outcomes)

Developers also should 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. Lastly, they should describe how the 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 for Evaluating the 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* in relation to the measure's specifications. Consideration of reliability and validity can help to address the following questions:

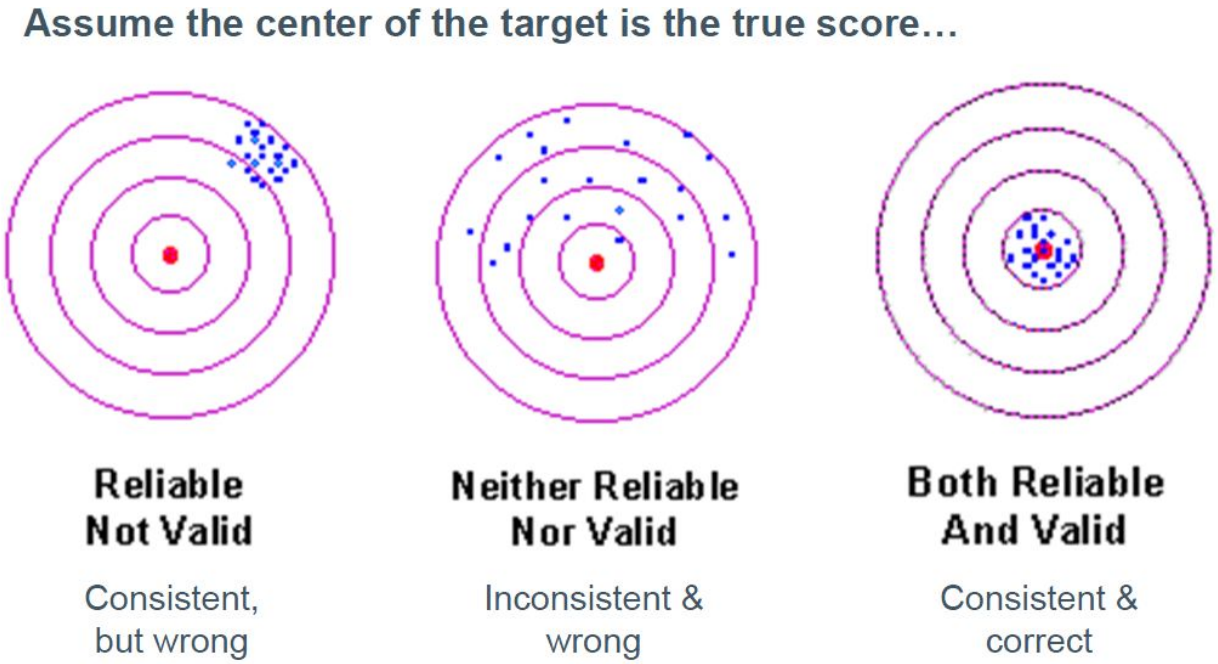
- Are the specifications clear so that everyone will calculate the measure in the same way?
- Is the variation between providers primarily due to real differences? Or is it because there is a lot of "noise" in the measurement?
- Is the measure actually measuring what it is intended to measure (i.e., 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 do not 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. As a result, for some measure types, validity testing can also be used to demonstrate the reliability of the measure.

**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 (1) how the measure is constructed—i.e., are the specifications written so that the measure can be computed consistently and (2) 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 patient/encounter level (formerly known as data element level) or at the accountable entity level (formerly known as performance measure score level) 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" patient/encounter level validity testing in place of patient/encounter level reliability testing
- for new measures, present evidence of the face validity of the accountable entity level measure (i.e., performance measure score) as an indicator of quality rather than conduct empirical

validation (empirical validity testing is expected for previously endorsed measures that are undergoing maintenance evaluation)

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 and standardize standing committee evaluation of measure reliability and validity (see Algorithms 2 and 3).

**Key Points for Evaluating Scientific Acceptability**

- Scientifically acceptable measures must be both reliable and valid.
- Empirical demonstration of reliability and validity is expected, although for new measures, demonstration of face validity of the measure score 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 testing results.
- Reliability and validity must be demonstrated for the measure as specified (including data source and level of analysis).
- NQF allows testing at either the patient/encounter level (formerly known as data element level) or at the accountable entity level (formerly known as 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.
- All three subcriteria under Scientific Acceptability are "must-pass"; therefore, each must be met in order to be recommended for endorsement.

### 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 patient/encounter level, the accountable entity level, or both; note, however, that patient/encounter level reliability testing is not required if patient/encounter level validity has been demonstrated. Testing at the patient/encounter 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 accountable entity level addresses the *precision* of the measure; such testing uses data that have been

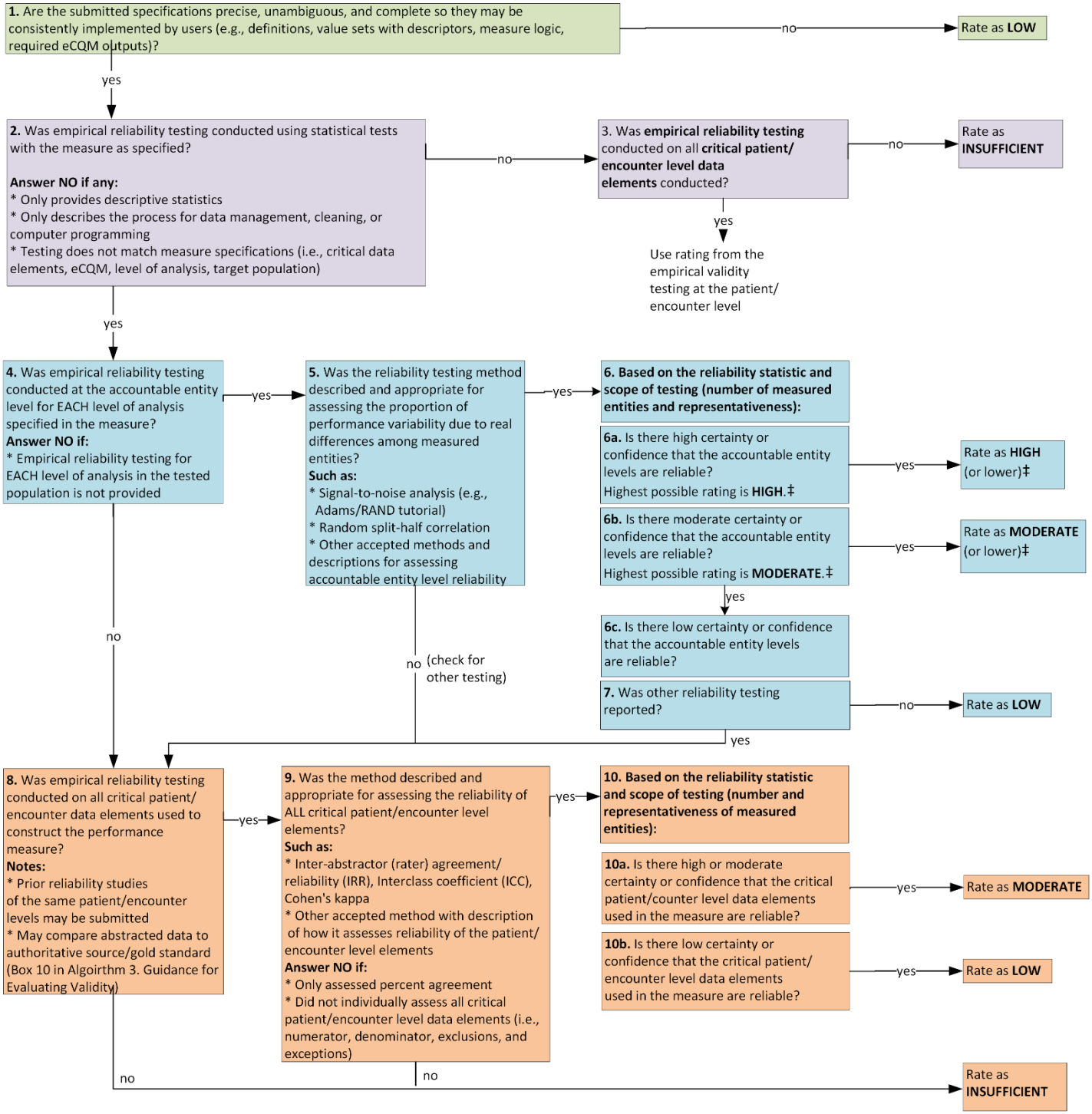
aggregated across providers. Developers also can choose from a variety of methods and statistics to test reliability. NQF is not prescriptive about the methods nor about the results; however, when evaluating the reliability of a measure, Standing Committees should consider the appropriateness of the 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 (For eCQMs, this includes following the industry accepted technical specifications for eCQMs and value sets)
> 2a2. Reliability testing—data elements or measure score

Guidance for evaluating reliability is provided in Algorithm 2.

## Algorithm 2. Guidance for Evaluating Reliability




‡ This is the highest **overall** possible rating **for reliability testing**, but it may be lower, depending on the strength of the patient/encounter level **reliability** testing results. If patient/ encounter level **reliability** testing is provided, these results must also be considered.

### Key Points for Evaluating Reliability

- Reliability refers to the repeatability and precision of measurement.

- Measurement precision reflects the ability to distinguish difference between providers that are due to quality of care rather than chance.
- Precise specifications provide the foundation for achieving consistency in measurement. All data elements must be clearly defined.
- Requirements for eCQM specifications include us of The Health Quality Measures Format (HQMF); The Quality Data Model (QDM); The Clinical Quality Language (CQL) and published value sets withing 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)
- Patient/encounter 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)
    - Common method is inter-rater reliability (common statistics include kappa; intra-class correlation coefficient)
    - Not required if patient/encounter validity is demonstrated

- Accountable entity reliability
    - Addresses the precision of the measure
    - Uses data that have been aggregated across providers
    - Common methods is signal-to-noise analysis

- When evaluating empirical validity testing of the measure, consider whether:
    - 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)

**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 (i.e., a higher score on a quality measure reflects higher quality of care).

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 at initial endorsement. Empirical testing is expected for maintenance of endorsement. As with reliability testing, developers can choose from a variety of methods and statistics to test validity empirically.

Although various terms 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) related to the quality construct 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 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 an explanation of 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 patient/encounter level, the accountable-entity level, or both. Testing at the patient/encounter 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 accountable entity 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 accumulate 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, or use of multiple data sources or methods that result in different scores and conclusions about quality. Other threats to validity may include systematically 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. Validity testing
2b2. Justification of exclusions
2b3. Risk adjustment (for outcome and resource use measures, possibly others)
2b4. Identification of differences in performance
2b5. Comparability of data sources/methods
2b6. Missing data

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, not associated with quality of care provided, and that could influence patient outcomes or resource use. The purpose of risk adjustment is to "level the playing field" so that, when

patients or other stakeholders compare providers, the differences in performance scores are due to true differences in the quality of care that is provided rather than to differences in patient groups (e.g., one provider's patients may be sicker than those of another provider but receive the same quality of care).

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). Currently, NQF's policy regarding risk adjustment does not prohibit including social risk factors in the risk adjustment strategy.

Specifically, for outcome measures and other measures as appropriate (e.g., cost/resource use), an evidence-based, risk adjustment strategy (e.g., risk models, risk stratification) should be specified. The risk adjustment method should consider patient factors (including clinical, functional, and social risk factors) that influence the measured outcome (but not factors related to disparities in care or the quality of care), are present at start of care, and demonstrate adequate discrimination and calibration, or alternatively, a rationale or data to support no risk adjustment/stratification should be provided.

When deciding which clinical, health status, and/or social risk factors are appropriate for a risk adjustment strategy, Standing Committees should consider the following:

- Clinical/conceptual relationship with the outcome of interest
- Empirical association with the outcome of interest
- Variation in prevalence of the factor across the measured entities
- Present at the start of care
- Is not an indicator or characteristic of the care provided (e.g., treatments, expertise of staff)
- Resistant to manipulation or gaming
- Accurate data that can be reliably and feasibly captured
- Contribution of unique variation in the outcome (i.e., not redundant)
- Potentially, improvement of the risk model (e.g., risk model metrics of discrimination, calibration)
- Potentially, face validity and acceptability

To aid Standing Committees as they consider inclusion (or not) of social risk factors in a risk adjustment strategy, NQF has directed that the following be included as part of the measure submission when there is a conceptual relationship and evidence that sociodemographic factors affect an outcome or process of care reflected in a performance measure:

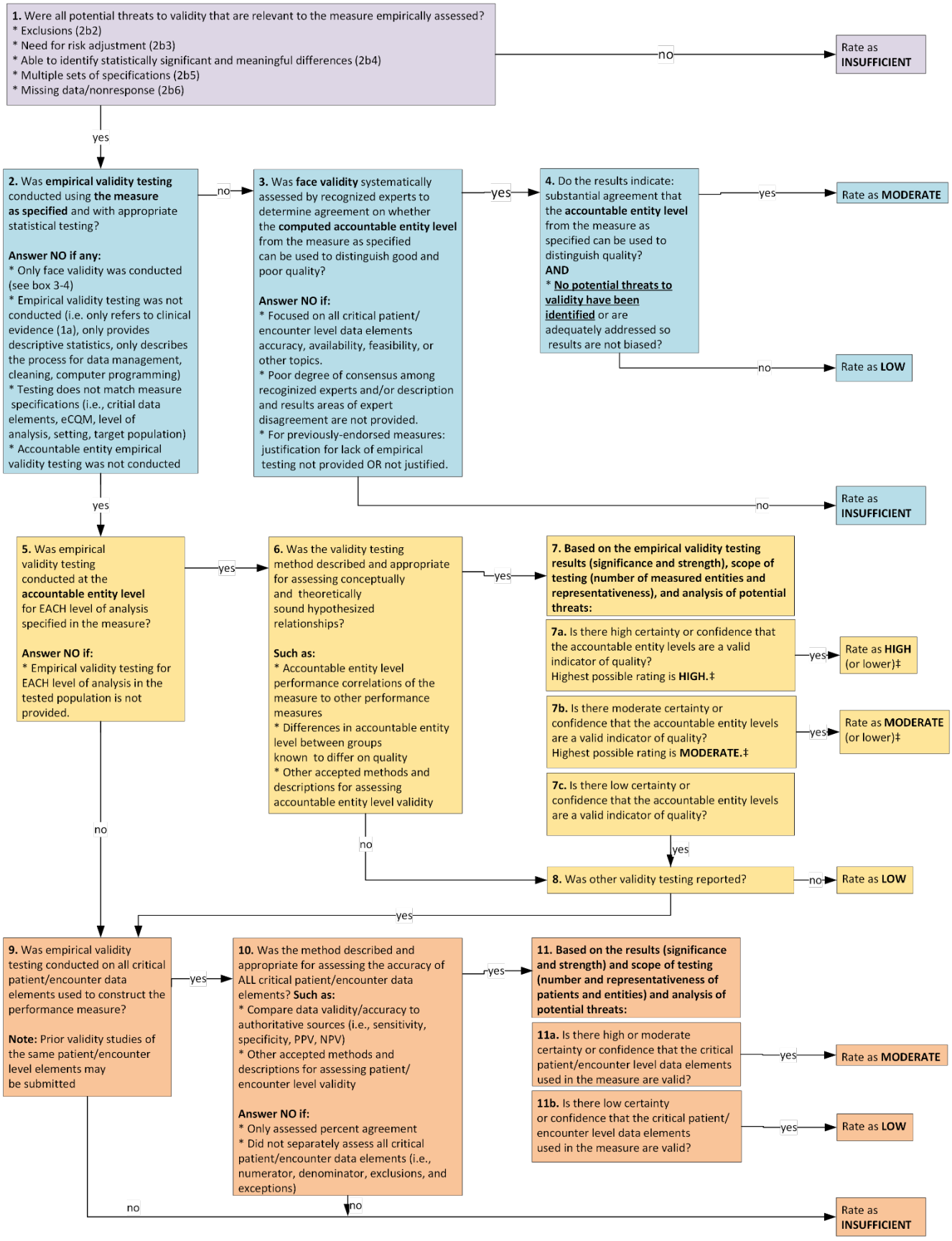
- A detailed discussion of the rationale and decisions for selecting or not selecting social risk factors and methods of adjustment, including a conceptual description of relationship to the outcome or process; empirical analyses; and impacts of limitations of available social risk data and/or potential proxy data)

Guidance for evaluating validity is provided in Algorithm 3.

**Key Points for Evaluating Validity**

- Validity refers to the correctness of measurement: that the measure is, in fact, measuring what it intends to measure and that the results of the measurement allow users to make the right conclusions.
- Testing should be done for the measure as specified (including data source and level of analysis).
- Patient/encounter 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 and specificity)
- Accountable entity 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 the following:
        - 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); and
    - potential threats to validity are adequately assessed and accounted for.
- Face validity—the subjective determination that, on the face of it, a measure appears to reflect quality of care—is the weakest demonstration of validity but is accepted by NQF for new measures as an indicator of quality (if systematically assessed); method and results should be described (more than simple statement that a measure was considered to be valid).

## Algorithm 3. Guidance for Evaluating Validity




‡ This is the highest **overall** possible rating for **validity testing**, but it may be lower, depending on the strength of the patient/encounter level validity testing results. If patient/ encounter level validity testing is provided, these results must also be considered.

**Example**

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

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

While subcriterion 1c 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 fit 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, this also is a "must-pass" subcriterion. NQF is not prescriptive about the methods used in the analyses that address this subcriterion: the methods used should follow from the quality construct that is described in subcriterion 1c.

**Key Points for Evaluating 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 eCQMs hinges on the data elements that are included in the measure and the logic that is used to compute the measure. Thus, for eCQMs, a summary of a feasibility assessment is required. Ideally, developers would utilize a standard scorecard to reflect this summary. 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 an identified feasibility issue; and a rationale and plan for addressing the feasibility issues. For eCQMs, feasibility of the measure logic should include unit testing of the measure logic using a simulated data set, such as the testing provided by the CMS BONNIE tool. Measure logic assessment demonstrates that each branch of the measure logic can not only be processed but also performs as expected when processed by other eCQM reporting systems.

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

**Key Points for Evaluating 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 are available in electronic form (e.g., EHR or other electronic sources); and
  - the data collection strategy is ready to be put into operational use.
- A summary of a formal feasibility assessment should be provided for eCQMs.
- 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, and policymakers) are using or could use performance results for both accountability and performance improvement to achieve the goal of high quality and 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. This fourth major criterion comprises of two subcriteria (Use and Usability), which are evaluated separately.

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, NQF expects measures to be in use in accountability programs by the time of endorsement maintenance and be publicly reported within six years of initial endorsement. The Use subcriterion goes beyond simply requiring that measures are being used. This subcriterion also addresses the growing desire to allow those being measured (or others) to see measure results and data, receive assistance in interpreting the results and data, and provide feedback on the measure and its implementation. These two facets of Use are "must-pass" requirements for maintenance of endorsement.

The Usability subcriterion under Usability and Use requires that use of the measure has led to demonstrable improvement in healthcare quality. Evidence of improvement could include improvement in measure performance over time, an increase in the number of individuals who are receiving high quality care over time, or the demonstration of a causal link between improvement activities associated with a particular measure and a desired outcome(s). Lastly, the Usability subcriterion 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 the Usability subcriterion is provided in Table 2. The rating options for the Use subcriterion are either Pass or No Pass.

**Key Points for Evaluating Usability and Use**

- Measures are not required to be in use at initial endorsement but 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.
- NQF is not prescriptive about the scope and processes used in providing and obtaining feedback for a measure; nonetheless, providing and obtaining feedback on the measure is expected by the time of measure maintenance.
- 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;
  - feedback on the measure has been provided by those being measured or other users;
  - the performance results have been used to further the goal of high quality, efficient healthcare; and
  - the benefits of the measure outweigh any potential unintended consequences.
- Usability and Use are not must-pass criteria at initial endorsement; however, the subcriteria dealing with use of the measure (in accountability/public reporting programs and acquiring feedback on the measure) are must-pass for maintenance measures.

### *Criterion 5: Related and Competing Measures*

NQF endorses national consensus standards—and this implies parsimony and standardization to the extent possible. Duplicative measures and/or those with similar but not identical specifications increase measurement burden and 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 and 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 which case, it will recommend the superior measure as suitable for endorsement but would not recommend the competing measure(s). 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 for Evaluating 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