NATIONAL QUALITY FORUM

Scientific Methods Panel Orientation Meeting: NQF Evaluation Criteria Tutorial

NQF Methods Panel Team

August 28, 2019

Welcome, Introductions, and Roll Call

NQF's Scientific Methods Panel Team

Michael Abrams
Senior Director

Karen Johnson
Senior Director

Andrew Lyzenga
Senior Director

Roara Michael
Project Manager

Yetunde Ogungbemi
Project Manager

Sam Stolpe
Senior Director

Ashlie Wilbon
Senior Director

NQF Scientific Methods Panel (SMP) Team

- Content Leads
  - *Karen Johnson, Senior Director (Team Lead)*
  - *Ashlie Wilbon, Senior Director*
  - *Andrew Lyzenga, Senior Director*
  - *Sam Stolpe, Senior Director*
  - *Michael Abrams, Senior Director*

- Project Management
  - *Yetunde Ogungbemi, Project Manager*
  - *Roara Michael, Project Manager*

New SMP Members

Daniel Deutscher, PT, PhD National Director of Research and Development, Maccabi Healthcare Services
Joseph Hyder, MD Associate Professor, Mayo Clinic
Sean O'Brien, PhD Associate Professor of Biostatistics and Bioinformatics, Duke University Medical Center
Patrick Romano, MD, MPH Professor, University of California Davis
David Salkever, PhD Professor Emeritus, Johns Hopkins Bloomberg School of Public Health
Alex Sox-Harris, PhD, MS Associate Professor of Research, Department of Surgery, Stanford University
Terri Warholak, PhD, RPh, CPHQ, FAPhA Assistant Dean of Academic Affairs and Assessment and Professor at the University of Arizona, College of Pharmacy
Eric Weinhandl, PhD, MS Senior Director, Epidemiology and Biostatistics, Fresenius Medical Care North America

Current Scientific Methods Panel Members

J. Matt Austin, PhD	Zhenqiu Lin, PhD
Bijan Borah, MSc, PhD	Jack Needleman, PhD
John Bott, MBA, MSSW	David Nerenz, PhD, Co-chair
David Cella, PhD, Co-chair	Eugene Nuccio, PhD
Lacy Fabian, PhD	Jennifer Perloff, PhD
Marybeth Farquhar, PhD, MSN, RN	Sam Simon, PhD
Jeffrey Geppert, EdM, JD	Michael Stoto, PhD
Laurent Glance, MD	Christie Teigland, PhD
Sherrie Kaplan, PhD, MPH	Ronald Walters, MD, MBA, MHA, MS
Joseph Kunisch, PhD, RN-BC, CPHQ	Susan White, PhD, RHIA, CHDA
Paul Kurlansky, MD	

Agenda

- Opportunity for Questions from August 26th call
- Brief Overview of NQF's Evaluation Criteria
- Deeper Dive: Evaluating Reliability and Validity
- Key Points for Measure Evaluation
- Overview of Preliminary Analysis (PA) Form
- Advice from Inaugural SMP Members
- Opportunity for NQF Member and Public Comment
- Next Steps

Opportunity for Questions from August 26 Call

Measure Evaluation Criteria Overview

NQF Measure Evaluation Criteria for Endorsement

NQF endorses measures for accountability applications (public reporting, payment programs, accreditation, etc.) as well as quality improvement.

- Standardized evaluation criteria
- Criteria have evolved over time in response to stakeholder feedback
- The quality measurement enterprise is constantly growing and evolving—greater experience, lessons learned, expanding demands for measures—the criteria evolve to reflect the ongoing needs of stakeholders

Major Endorsement Criteria

- **Importance to measure and report**: Goal is to measure those aspects with greatest potential of driving improvements; if not important, the other criteria are less meaningful (**must-pass**)
- **Scientific acceptability of measure properties**: Goal is to make valid conclusions about quality; if not reliable and valid, there is risk of improper interpretation (**must-pass**)
- **Feasibility**: Goal is to, ideally, cause as little burden as possible; if not feasible, consider alternative approaches
- **Usability and Use** (**Use** is **must-pass** for maintenance measures): Goal is to use for decisions related to accountability and improvement;
- **Comparison to related or competing measures**

Criterion 1: Importance to Measure and Report

1. Importance to measure and report - 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.

1a. Evidence: *the measure focus is evidence-based*

1b. Opportunity for Improvement: *demonstration of quality problems and opportunity for improvement, i.e., data demonstrating considerable variation, or overall less-than-optimal performance, in the quality of care across providers; and/or disparities in care across population groups*

1c. Quality construct and rationale *(composite measures only)*

Criterion 2: Reliability and Validity – Scientific Acceptability of Measure Properties

Extent to which the measure, as specified, produces consistent (reliable) and credible (valid) results about the quality of health care delivery

2a. Reliability (must-pass)

2a1. Precise specifications including exclusions

2a2. Reliability testing—data elements or measure score

2b. Validity (must-pass)

2b1. Validity testing—data elements or measure score

2b2. Justification of exclusions—relates to evidence

2b3. Risk adjustment—typically for outcome/cost/resource use

2b4. Identification of differences in performance

2b5. Comparability of data sources/methods

2b6. Missing data

Criterion 3: Feasibility

Extent to which the required data are readily available, retrievable without undue burden, and can be implemented for performance measurement.

3a: Clinical data generated during care process

3b: Electronic sources

3c: Data collection strategy can be implemented

Criterion 4: Usability and Use

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

Use (4a) – must-pass for maintenance measures

4a1: Accountability and Transparency: *Performance results are used in at least one accountability application within three years after initial endorsement and are publicly reported within six years after initial endorsement.*

4a2: Feedback by those being measured or others: *Those being measured have been given results and assistance in interpreting results; those being measured and others have been given opportunity for feedback; the feedback has been considered by developers.*

Usability (4b)

4b1: Improvement: *Progress toward achieving the goal of high-quality, efficient healthcare for individuals or populations is demonstrated.*

4b2: Benefits outweigh the harms: *The benefits of the performance measure in facilitating progress toward achieving high-quality, efficient healthcare for individuals or populations outweigh evidence of unintended negative consequences to individuals or populations (if such evidence exists).*

Criterion 5: Related or Competing Measures

If a measure meets the four criteria and there are endorsed/new related measures (same measure focus or same target population) or competing measures (both the same measure focus and same target population), the measures are compared to address harmonization and/or selection of the best measure

- 5a. The measure specifications are harmonized with related measures **OR** the differences in specifications are justified
- 5b. The measure is superior to competing measures (e.g., is a more valid or efficient way to measure) **OR** multiple measures are justified

Questions?

A Deeper Dive: Evaluating Reliability and Validity

Key Definitions

- Level of analysis
  - *Term used to describe the "accountable entity" (i.e., the entity whose performance is being measured)*
- Provider
  - *Umbrella term that describes the entity whose performance is being measured*
- Maintenance measure
  - *Previously endorsed measure that is being evaluated for re-endorsement*
- Measure score
  - *The calculated results of a measure*
- Data elements
  - *Individual items of information used to calculate measures*

Some Current Assumptions

- There will always be some error in performance measurement
  - *Random error affects reliability; systematic error affects validity*
- Reliability and validity are **not static properties of a measure** (both can vary under conditions of implementation)
- Neither reliability nor validity is an all-or-none property and is instead a **matter of degree**
  - *Considerations are scope of testing, method used, and results obtained*
- Reliability does not guarantee validity, or vice-versa

Evaluating Scientific Acceptability – Key Points

- In general, empirical analysis is expected
  - *Face validity of the measure score is allowed for new measures, but not for maintenance measures unless there is justification*
- NQF is not prescriptive about how empirical measure testing is done
- NQF has not set minimum thresholds for reliability or validity testing results
- Reliability and Validity are “must-pass” criteria
- There may be different/additional testing requirements depending on measure type

Criterion 2a: Reliability

2a. **Reliability (must-pass)**

2a1. Precise specifications including exclusions

2a2. Reliability testing—data elements or measure score

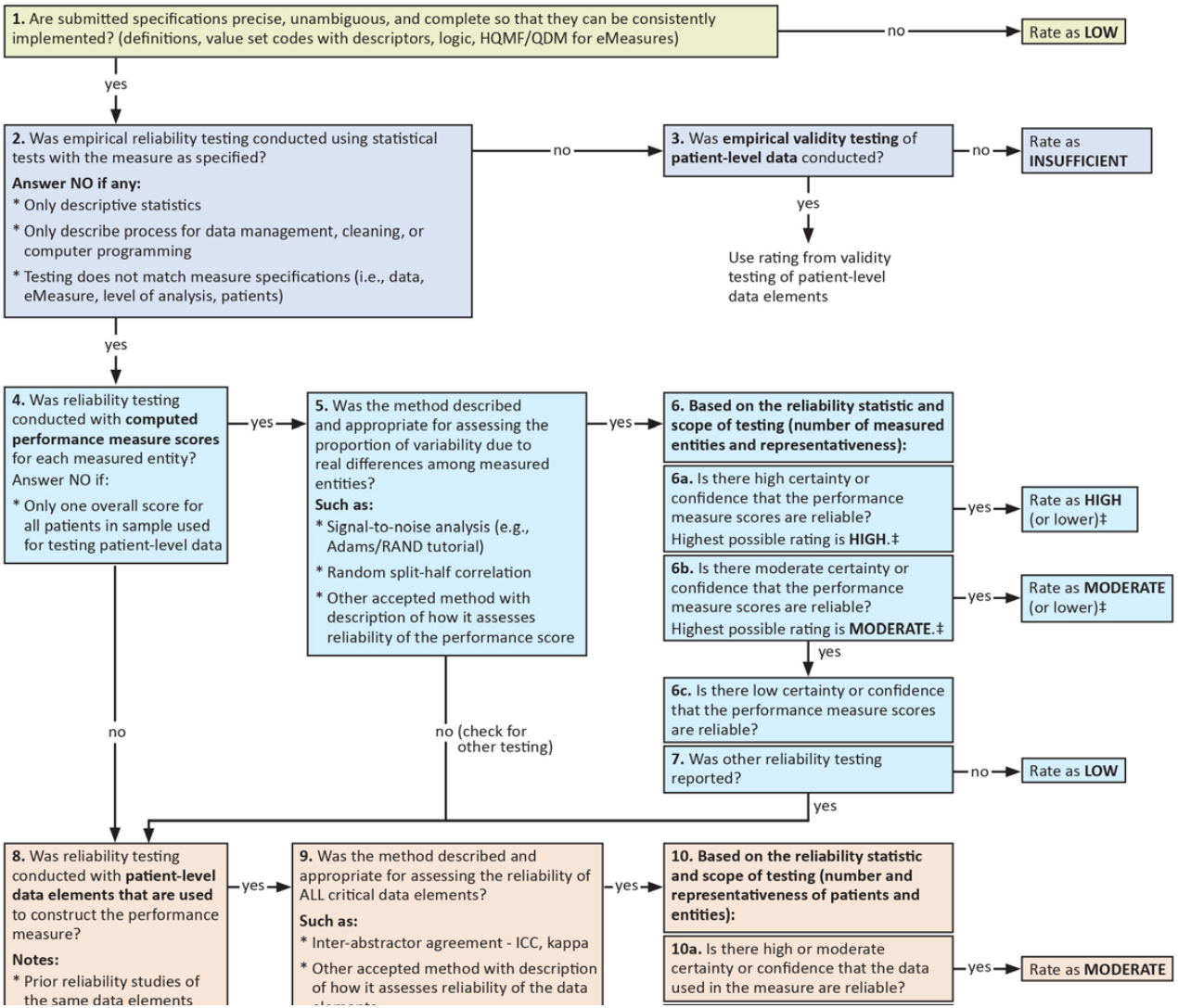
- Specifications
  - *Must be precise, unambiguous, complete*
  - *eCQM logic will be evaluated by NQF staff*
- Testing
  - *NQF distinguishes between testing of the data elements and testing of the measure score*
  - *Testing can be done on samples*
  - *Prior evidence may be used as appropriate*

Reliability Testing - Key Points

- Reliability of the ***measure score***: Applies to result that is aggregated to the specified level of analysis; quantifies precision/repeatability and risk of misclassification
  - *Example: Signal-to-noise analysis*
  - *Example: Split-half test*
- Reliability of the ***data elements:*** Applies to patient-level data used in a measure; quantifies the reproducibility of the data elements used in the measure and uses patient-level data
  - *Example: Inter-rater reliability*
  - *Example: Internal consistency (for multi-item scales)*
  - *At minimum, for numerator, denominator, exclusions*
- Consider whether testing used an appropriate method and included adequate representation of providers and patients and whether results are within acceptable norms

Rating Reliability: Algorithm #2

Algorithm 2. Guidance for Evaluating Reliability



Criterion 2b: Validity

2b. Validity (must-pass)

2b1. Validity testing—data elements or measure score

2b2. Justification of exclusions—relates to evidence

2b3. Risk adjustment—typically for outcome/cost/resource use measures

2b4. Identification of differences in performance

2b5. Comparability of data sources/methods

2b6. Missing data

Validity Testing – Key points

Empirical testing

- Measure score – assesses a hypothesized relationship of the measure results to some other concept; assesses the correctness of conclusions about quality
  - *We don't worry too much about labels such as concurrent, predictive, etc.*
  - *Often assessed via correlations (e.g., ICCs)*
  - *Adequacy of comparator(s) may require clinical judgement*
- Data element – assesses the correctness of the data elements compared to a "gold standard"
  - *We want sensitivity/specificity; have allowed less (kappa values showing agreement with gold standard)*

Validity Testing – Key Points

Face validity

- Subjective determination by experts that the measure appears to reflect quality of care
  - *Systematic and transparent process*
  - *Explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality*
  - *Degree of consensus and any areas of disagreement must be provided/discussed*
- We see content validity of instruments as something different than face validity "testing"
- Acceptable for new measures, but empirical testing is expected for measures up for re-endorsement, unless there is good justification for lack of empirical testing

Threats to Validity

Exclusions

- Any patients inappropriately excluded from measurement?
- Exclusions consistent with evidence, and of sufficient frequency to warrant inclusion?
- Clinical judgement often required, particularly around evidence

Threats to Validity

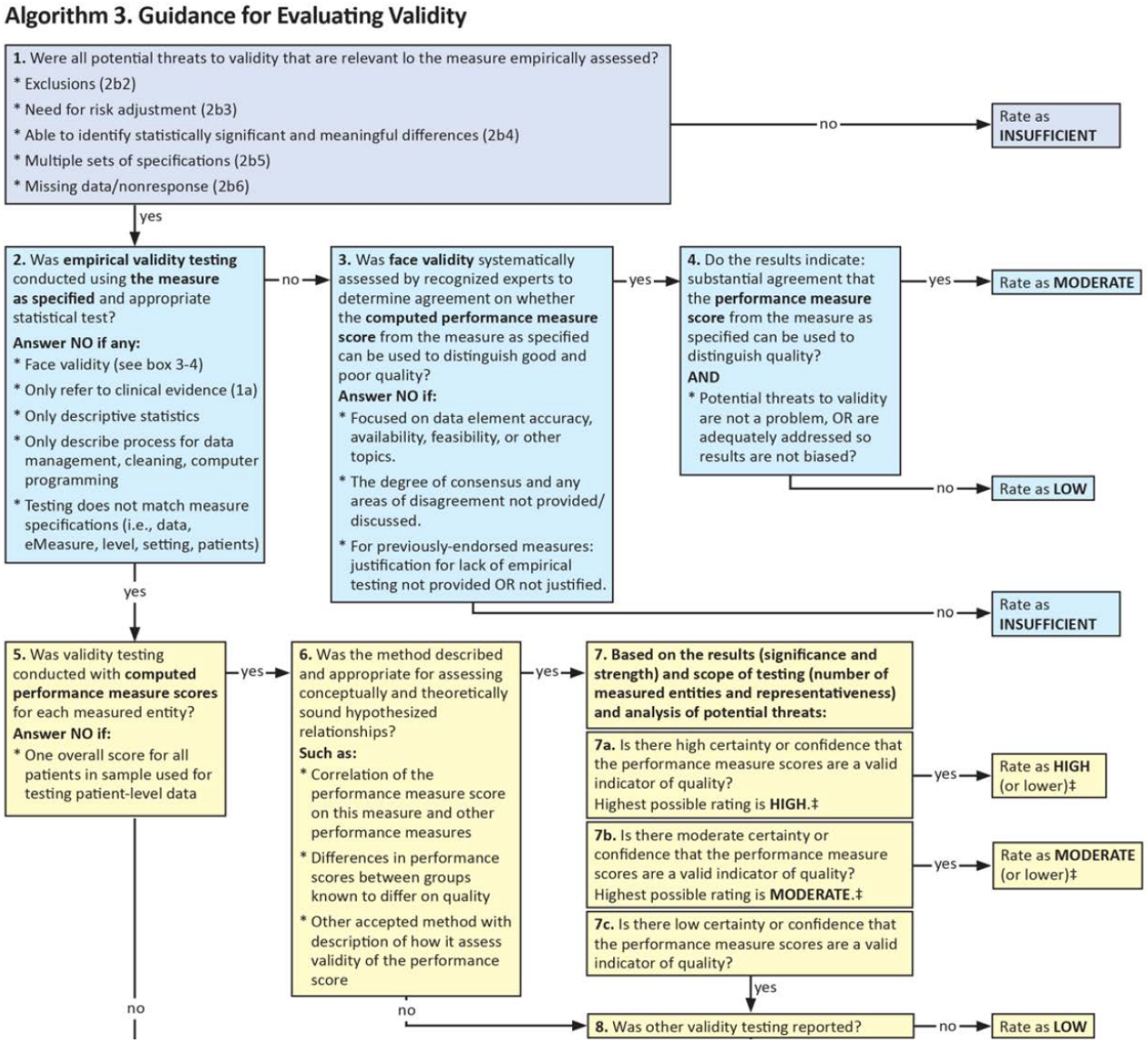
Risk-adjustment

- For outcome and C/RU measures, risk-adjustment is expected – but developer can provide rationale/data to support not adjusting
- Social risk factors can be included, if there is a conceptual rationale
- Conceptual rationale for risk factors should be included in submission materials
- Expect calibration/discrimination statistics, as well as analysis to support inclusion (or not) of social risk factors
- Cannot "fail" a measure due to inclusion (or not) of particular risk factors
  - *Concerns with discrimination, calibration, or overall method of adjustment are still grounds for rejecting a measure*
- Clinical judgement often required

Threats to Validity

- Meaningful differences in performance
  - *Look for statistical differences across providers*
    - » Often, statistical testing not conducted
  - *Also interested in clinically meaningful differences; therefore, clinical judgement required*
- Comparable results for measure scores that are generated with multiple data sources/methods
  - *Very seldom provided*
- Missing data do not produce biased results
  - *Ideally, we expect sensitivity analysis*
  - *Often get frequency analysis*
  - *May require clinical judgement*

Rating Validity: Algorithm #3



Questions?

Key Points for Measure Evaluation

Materials for Evaluation

- Submission information
  - *Specifications*
  - *Testing Attachment*
  - *Feasibility Scorecard (for eCQMs)*
  - *For maintenance measures, summary of the last evaluation*
    - » For informational purposes only; decisions made in last evaluation should not influence your evaluation
- If applicable, previous SMP evaluation materials
- Preliminary Analysis Form
  - *You will complete one for each assigned measure (informal discussions between SMP members are allowed)*
- Resources available to you
  - *2018 Criteria and Guidance Document*
  - *Standing Committee Guidebook (section 7)*
  - *"Key Points" guidance document*
  - *Methods Panel staff (for questions about the criteria)*

Reminders: A Few Things to Check

- Measures should be tested as specified
  - *Level of analysis (if multiple LoAs specified, each should be tested separately)*
  - *Data source (separate testing likely needed, but maybe not always)*
  - *Care setting (separate testing preferred, but at minimum, testing dataset should include data from all specified settings)*
  - *It is possible for you to "pass" part of the measure*
- Often there are several performance measures included under one NQF number
  - *Each must be evaluated separately; some might pass and others not pass*
- Look for/point out any inconsistencies in submissions

Overall Ratings

- High
  - *Score-level testing is required*
  - *A measure may be eligible for "HIGH" but the sampling method/results may make you choose "MODERATE" instead*
- Moderate
  - *The highest eligible rating if only data element testing or face validity testing is conducted*
  - *A measure may be eligible for "MODERATE" but the sampling method/results may make you choose "LOW" instead*
- Low
  - *Used primarily if testing results are not satisfactory*
- Insufficient
  - *Use when you don't have sufficient information to assign a HIGH, MODERATE, or LOW rating*
    - » Example: unclear specifications; unclear testing methodology

Differences in Testing Requirements by Measure Type

- Health outcomes, intermediate clinical outcomes, cost/resource use, structure, process
  - *For both reliability and validity, NQF requires EITHER data element testing OR score-level testing*
    - » We'd prefer both, but currently do not require both
    - » Impacts rating, as described above
    - » Exception: face validity for new measures accepted
  - *If data element validity testing provided, we **do not** require additional reliability testing*
    - » In this case, use the rating you give for validity as the rating for reliability
    - » This is not as common as it used to be
    - » If in doubt, contact staff!!

Differences in Testing Requirements by Measure Type

Composite measures

- NQF has specific definitions for "composite" measures
  - *"Traditional" composites*
  - *All-or-none measures*
  - *Does NOT include multi-item scales in surveys/questionnaires*
- Require reliability testing of the composite measure score
  - *Can also show reliability testing of the components, but this is not sufficient to pass the criterion*
- Score-level validity testing not required until maintenance
- Additional subcriterion: Empirical analyses to support the composite construction
  - *How this is addressed by the developer will depend on the type of composite*

Differences in Testing Requirements by Measure Type

Instrument-based measures

- For reliability and validity, require testing at **both** levels
  - *Data element level → must demonstrate R/V of the relevant items in the instrument*
  - *Measure score level → testing of the actual performance measure*
- We do allow multiple performance measures under same NQF number: need only one form, but may require multiple ratings

Differences in Testing Requirements by Measure Type

eCQMs (eMeasures):

- Testing from 2 EHR systems required
  - *While more would be great, it is not required*
- Reliability testing not required if based on data from structured data fields. Unstructured fields require both reliability and validity testing
- New (summer 2019) requirement: data element validation
- We will also provide feasibility scorecard

One "Exception" to our Testing Requirements

For some measure types, if data element validity testing is provided, we **do not** require additional reliability testing

- In this case, use the rating you give for validity as the rating for reliability
- This is not as common as it used to be.
- If in doubt, contact staff!!

Questions?

Preliminary Analysis Form

Preliminary Analysis Form

- Includes instructions, useful links, and testing requirements by measure type
- Includes notes about where information should be included on submission forms
- Form designed for ease of use – but feel free to write as much explanation as you want

Your (de-identified) responses will be made public!

Questions?

Advice from Inaugural SMP Members

Opportunity for Member and Public Comment

Important Dates

- Measure-specific DOIs: Due **August 23**
- SMP review of measures: **September 3-27**
- SMP in-person meeting: **October 28-29**
  - *Travel information will be distributed ~1 month prior to meeting*

- Have questions? Contact us at:
  - *MethodsPanel@qualityforum.org*

Questions?

THANK YOU