NATIONAL QUALITY FORUM
Driving measurable health improvements together

Meeting Summary

Scientific Methods Panel June 2022 Advisory Web Meeting

The National Quality Forum (NQF) convened the Scientific Methods Panel (SMP) for a web meeting on June 14, 2022, to discuss possible improvements to the NQF measure evaluation criteria.

Welcome, Roll Call, and Review of Meeting Objectives

Dr. Matthew Pickering, NQF senior director, welcomed the SMP and participants to the web meeting. Dr. Pickering and SMP Co-Chairs Drs. David Nerenz and Christie Teigland provided opening remarks and reviewed the following meeting objectives: (1) review and discuss the preliminary results of the impact analysis, (2) discuss and provide further guidance on divergent testing results for the patient/encounter and accountable-entity levels, (3) review and consider face validity testing requirements and its acceptability for maintenance endorsement, and (4) review and consider updates to the ratings assigned to the scientific acceptability criteria.

Review and Discuss the Preliminary Results of the Impact Analysis

In December 2021, the SMP finalized three recommendations for improvements to NQF's criteria:

(1) Establish reliability testing thresholds, which measures must meet at initial endorsement and maintenance endorsement
(2) Require accountable entity-level reliability testing at maintenance endorsement
(3) Require accountable entity-level empirical validity testing at maintenance endorsement (i.e., not accepting face validity for maintenance review nor patient/encounter-level empirical testing only)

To garner additional input as to whether these recommendations should be implemented in NQF's measure evaluation criteria and guidance, NQF staff analyzed all endorsed measures reviewed by the SMP to understand which measures currently meet or do not meet these recommendations. A total of 163 measures have been reviewed by the SMP since its inception in 2017. However, NQF staff did not include measures that are currently under endorsement review. This includes new measures within the fall 2021 and spring 2022 review cycles. This limited the impact analysis pool to 138 measures. In addition, NQF staff also stratified the results by measures that are in use within federal programs.

Of the 138 measures:

- Patient/Encounter Testing Category
  - 81 measures did not undergo patient/encounter-level reliability testing; and
  - 57 measures did undergo patient/encounter-level reliability testing.
    - Of the 57 measures:
      - 27 met the recommended reliability thresholds;
      - seven did not meet the recommended thresholds; and
      - there were 23 measures for which some testing met and did not meet the threshold, or the threshold could not be confirmed.
        - For example, NQF did not include the measures for which some data elements' testing results met and did not meet the

threshold. Based on the information in the table, it was unclear whether the overall measure should be classified as meeting or not meeting the threshold.

- Accountable Entity Testing Category
    - 11 measures did not undergo accountable entity-level reliability testing; and
    - 127 measures did undergo accountable entity-level reliability testing.
        - Of the 127 measures:
            - 89 met the recommended thresholds;
            - 13 did not meet the recommended thresholds; and
            - there were 25 measures for which some testing met and did not meet the threshold, or the threshold could not be confirmed.
                - For example, NQF did not include the measures for which the threshold could not be confirmed because the methods presented were not included in the information in the table. Therefore, it was unclear whether the measure should be classified as meeting or not meeting the threshold.
- Empirical Validity Testing Category
    - 117 measures did undergo accountable entity-level empirical validity testing; and
    - 21 measures did not undergo accountable entity-level empirical validity testing.

Of the 138 measures, 59 are used within federal programs. Of the 59 measures:

- Patient/Encounter Testing Category
    - 33 measures did not undergo patient/encounter-level reliability testing; and
    - 26 measures did undergo patient/encounter-level reliability testing.
        - Of the 26 measures:
            - 10 met the recommended thresholds;
            - five did not meet the recommended thresholds; and
            - there were 11 measures for which some testing met and did not meet the threshold or the threshold could not be confirmed.
                - As stated above, NQF did not include these measures in the final counts because, based on the information in the table, it was unclear whether the measure should be classified as meeting or not meeting the threshold.
- Accountable Entity Testing Category
    - four measures did not undergo accountable entity-level reliability testing; and
    - 55 measures did undergo accountable entity-level reliability testing.
        - Of the 55 measures:
            - 34 met the recommended thresholds;
            - 10 did not meet the recommended thresholds; and
            - there were 11 measures for which some testing met and did not meet the threshold, or the threshold could not be confirmed.
                - As stated above, NQF did not include these measures in the final counts because, based on the information in the table, it was unclear whether the measure should be classified as meeting or not meeting the threshold.
- Empirical Validity Testing Category
    - 50 measures did undergo accountable entity-level empirical validity testing; and
    - nine measures did not undergo accountable entity-level empirical validity testing.

After reviewing the results of the impact analysis, NQF posed the following discussion questions to the SMP:

(1) If a measure uses multiple methods, how would we classify whether a measure meets the threshold when one method for establishing reliability is good and when one is not as good and vice versa?
(2) There are cases in which the overall/mean reliability is high, yet there is a nontrivial percentage of accountable entities for which the reliability is below the threshold. How would the SMP envision applying thresholds to cases such as this one?
(3) There are some measures for which the reliability testing was deemed satisfactory by the SMP, yet it does not clearly fall into the methods outlined in the thresholds table. How would the SMP envision applying thresholds to cases such as this one?

An SMP member asked for clarification on NQF's methods, specifically whether the measures reviewed were further broken out by initial versus maintenance endorsement. NQF staff noted that because all measures reviewed are NQF-endorsed, they are measures that would return for maintenance endorsement.

The SMP agreed that the results of the impact analysis aligned with its expectations. Furthermore, the SMP members were not surprised by the list of measures that did not meet the thresholds due to the nature of what they are measuring and how they are constructed. For example, the list included many measures of infrequent events, which tend to have varying distributions across facilities and can impact reliability estimates. While the SMP was in general agreement about the results, one SMP member did note that for the accountable entity-level results, approximately two-thirds of the measures would pass if the reliability thresholds were applied, which is a small proportion in their opinion. Dr. Pickering stated that while one-third of the measures not passing the thresholds at the accountable-entity level is indeed a large number, NQF believes that the thresholds may not have a significant impact on NQF's portfolio of measures, generally speaking. Dr. Pickering continued by stating that for measures that do not meet the thresholds, the SMP may need to evaluate additional considerations about the measures, such as a developer's rationale for not meeting the threshold, and ultimately pass the measure. Additionally, NQF Chief Scientific Officer Dr. Elizabeth Drye noted that the most significant impact is seen with the measures in use within federal programs and that because of this, the conversation about implementing the thresholds would need to involve more stakeholders. One SMP member responded by stating that this might be the exact reason to implement the thresholds and recommendations, further stating that measured entities in the real world are facing the impact of these measures. They also stated that if the measures in use are not methodologically sound, perhaps they need to be revisited to ensure the programs are truly reflecting quality of care.

Regarding the question of how the SMP envisions applying the thresholds when the developer presents an overall or mean reliability that is high despite the distribution showing a nontrivial number of accountable entities that do not meet the threshold, the SMP largely discussed how to improve reliability results. For example, if the measure in question is a hospital-level measure and the developer specifies a minimum threshold of 50 cases in order to meet the reliability thresholds, then the solution would be to require and fully disclose a minimum sample size. Other SMP members cautioned against requiring a minimum sample size, as there are other methods that the developer could use to improve reliability, such as increasing the number of years of data collected or changing the construction of the measure to improve its intrinsic reliability. Because the SMP is not necessarily recommending that measures that do not meet the thresholds can never obtain endorsement, another SMP member suggested that the developer should provide justification for how they considered low reliability. For

example, developers could offer a persuasive response or perspective on these situations to explain, despite the low empirical reliability, how the measure is preferrable to another measure construction that would result in higher reliability but is not meaningful and therefore not preferrable for improving quality of care. Additionally, SMP members reiterated that NQF requires information about the distribution of reliability and that the developer should provide a minimum case volume if known; they also reiterated that the SMP needs to consider the measure exclusions and how a minimum case volume would impact those being included in the measure.

The SMP also considered situations in which measures were recommended for endorsement and therefore had reliability testing that the SMP deemed satisfactory; however, the testing did not clearly fall into the methods outlined in the thresholds table. NQF staff requested additional discussion of how the SMP would envision applying the thresholds. An SMP member noted that when different methods are used than those strictly laid out in the reliability thresholds table, the SMP needs to consider whether the testing method is evaluating the same concept as those outlined in the table or whether it is evaluating a completely different concept. If the methods are evaluating a different concept of reliability, then they cannot be evaluated against the threshold(s) in the table and must be considered on a case-by-case basis.

Additionally, the SMP discussed how the SMP envisions applying the thresholds when a measure employs multiple methods, and one method meets the threshold while the other does not. Many SMP members stated that having at least one appropriate method that meets the standard might be good enough to pass the measure. One SMP member noted that inherently when a measure developer calculates a reliability value that falls below the threshold, they will find another method to show better reliability. The SMP member emphasized that these instances are precisely why the SMP exists: to evaluate which findings are more relevant, determine why the results are divergent, and base its evaluation on the appropriate testing. Another SMP member suggested that in the example of inter-unit reliability (IUR) and profile inter-unit reliability (PIUR), the reliability decisions should be made based on the use of the measure, as IUR establishes reliability of performance scores across a distribution of accountable entities, whereas PIUR is best suited for determining whether the measure can reliably identify outliers. The points in this conversation relate directly to the discussion on the divergent testing results below.

Discuss and Provide Further Guidance on Divergent Testing Results for the Patient/Encounter and Accountable-Entity Levels

In the past, the SMP has expressed the need for additional guidance to assess how to evaluate when a measure developer submits multiple levels of testing when the testing results are divergent (i.e., one result is sufficient, but a second result is not). NQF criteria currently state that decisions regarding whether a measure's testing results are satisfactory are up to the discretion of the NQF-convened body reviewing the measure. The criteria also do not provide prescriptive guidance for how NQF-convened bodies should consider multiple testing results that are divergent. NQF staff asked the SMP to provide its suggestions for improvements or recommended changes to the NQF criteria and guidance. Additional details and caveats are outlined below.

In general, SMP members agreed that the SMP should give higher value to the results of the accountable entity-level testing, as the measure is always being used at the accountable-entity level. One SMP member further stated that if divergence occurs, developers should discuss their understanding on why the divergence occurred. The SMP can use this explanation in its discussion of whether the methodology is suitable. Generally, the SMP broadly agreed with this concept, with some members noting that they would go so far as to suggest that NQF should stop asking for patient/encounter-reliability testing, as

patient/encounter-reliability issues are only important insofar as they cause issues for accountable entity-level reliability testing. Dr. Pickering asked to discuss a situation in which the developer submits both levels of testing in which one meets the threshold while the other does not. One SMP member noted that if the testing does not meet the recommended criteria, which requires accountable entity-level testing at maintenance, then it would not pass. The SMP member then mentioned that only a subset of measures is subjected to patient/encounter-level reliability testing, and therefore, this conflict might occur less commonly in practice. When it does occur, the SMP can evaluate and interpret the results within the context of the different concepts they are testing.

However, another SMP member noted that this may have the unintended consequence of incentivizing developers to present less testing. For example, if a developer finds poor patient/encounter-level testing and this is not required, then it will be ignored in the measure evaluation process. The SMP suggested that developers should be encouraged to explain the divergent testing rather than eliminate it from their submission and that examples could be included in NQF's measure evaluation guidance and criteria. While most of the SMP agreed that accountable entity-level testing would take precedence over patient/encounter-level testing, one member noted that they would not consider accountable entity-level testing as primary for patient-reported outcome measures (PROMs) because the individual items in the survey instrument should be testing for both reliability and validity at the patient/encounter level. Therefore, it would be inappropriate for the SMP to evaluate a PROM under the assumption that accountable entity-level testing takes precedence over the patient/encounter level. Both patient/encounter and accountable-entity levels of testing are required for instrument-based measures.

One SMP member noted that it is not as clear what to do when accountable entity-level empirical validity testing results diverge from face validity results. Because many methods for empirical validity testing are underdeveloped and can produce unsatisfying results, there may be occasions in which excellent face validity trumps poor accountable entity-level empirical validity testing. Other SMP members noted that face validity relies on the person asked and the process used to establish face validity. As a result, empirical evidence should be presented and take precedence over face validity because without empirical evidence, it is difficult to know whether the measure is actually measuring quality in the sense that the developer asserts. Lastly, additional guidance on the specific requirements of face validity testing may be helpful to developers. Dr. Drye did note that when developers are creating measures in a new space and do not have a gold standard measure with which to compare it, face validity is often the only option. Dr. Drye also noted that with regard to risk adjustment, face validity is an incredibly important part of the model's validity. Face validity testing is often a structured process that involves convening experts and other stakeholder groups to assess a measure's validity and is not a casual endeavor. NQF's measure evaluation guidance and criteria require that any disagreement be noted and explained in the submission. One SMP member stated that while face validity is necessary, it may not be sufficient to show validity when the measure is ultimately going to be used empirically.

Review and Consider Face Validity Testing Requirements and Acceptability for Maintenance Endorsement

Currently, NQF's criteria allow a measure developer to submit face validity testing for new measures. There are specific requirements that the face validity testing must meet, as outlined in NQF's criteria. The process of collecting data must be systematic and transparent, gathered from identified experts, and must explicitly ask experts whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality of care. The degree of consensus and any areas of disagreement must be provided and/or discussed. Additionally, empirical validity testing is expected at the time of maintenance endorsement. If this is not possible, justification is required. In the past, the

SMP has expressed a desire to improve this policy by limiting the acceptability of face validity testing at maintenance since empirical validity testing results should be available by then. NQF staff asked the SMP to discuss the benefits and pain points of this policy, suggestions for improvement to the policy, and how the criteria can be bolstered.

One SMP member noted that although they agree that requiring empirical validity testing at maintenance is preferrable, they reiterated that the results that are presented are often unsatisfying. Furthermore, if by requiring empirical validity testing at maintenance the SMP continues to see the same low quality of results, it may not be worthwhile. The SMP has an opportunity to expand on what is meant by empirical validity and what are acceptable methods to establish empirical validity at the accountable-entity level.

Another SMP member noted that as a whole, the SMP should have more discussions on what an acceptable threshold for face validity looks like. The SMP member stated that a technical expert panel that supports a measure by a bare majority may not be an acceptable threshold, further noting that if the SMP is going to rely on face validity, stricter standards be in place.

The SMP focused the rest of conversation on acceptable justifications if a developer submitted face validity in lieu of accountable-entity empirical validity testing at maintenance endorsement. As noted above, there are several instances in which accountable-entity empirical validity testing may not be possible. These instances include the following: (1) measures of rare or serious reportable events because despite their low frequency, the general consensus is that they reflect breakdowns in quality of care and (2) patient experience measures because the consensus is that the concept of improving the patient experience is so important that they do not need to be empirically tested against other quality measures. However, the SMP noted that these are just two possibilities. Dr. Nerenz suggested that the SMP produce a list of examples that the SMP might be willing to accept as exception conditions for when empirical validity data are unnecessary. This would be helpful to developers to understand the types of justifications that the SMP might accept and could be integrated into NQF's measure evaluation criteria and guidance.

Review and Consider Updates to the Ratings Assigned to the Scientific Acceptability Criteria

At the SMP's last advisory meeting, the SMP suggested that NQF reconsider the ratings assigned to the scientific acceptability criteria. Currently, NQF's criteria use a four-part rating scale, which includes "high," "moderate," "low," and "insufficient." At a prior meeting, SMP members presented an alternative rating scale with options of "pass," "does not pass," and "insufficient." NQF staff asked the SMP to discuss whether these changes would better communicate results to measure developers, Standing Committees, and the public.

One SMP member started the conversation by noting that the reason the ratings of "high," "moderate," "low," and "insufficient" were initially used by the SMP was to distinguish between a measure's use and how scientifically acceptable the measures were based on that use. For example, measures being evaluated with the intention of being used in pay-for-performance programs would be required to have a high rating for reliability and validity. However, many SMP members noted that this is not how the ratings are applied. The majority of the SMP agreed that the high and moderate ratings have no discriminatory power, considering they both result in passing the measure.

In reference to the potential future state of the ratings presented by NQF staff, one SMP member noted that it seems that "does not pass" and "insufficient" result in the same outcome, so it is important to

clearly articulate their nuances. Dr. Pickering clarified that "insufficient" is used when there is not enough information for the convening body to make a determination about the results of the measure's testing. Additionally, Dr. Pickering noted that "does not pass" would be used when there is enough information, but based on NQF criteria, the measure does not pass.

Another SMP member noted that it might be more useful to change "insufficient" to "needs more information" so that it more clearly articulates the rating's meaning. However, SMP members agreed with updating the ratings from "high," "moderate," "low," and "insufficient" to new categories akin to "pass," "does not pass," and "insufficient."

Final Comments

Dr. Pickering asked the SMP members whether they had any additional final comments or thoughts about the discussions that were held during the meeting. One SMP member noted that NQF should consider tiering measures because some of the issues presented during the meeting could be implemented differently depending on the measure being reviewed. The SMP member referred to the prior conversation on empirical evidence as an example, noting that measures that have been in the field for a long time should, in theory, be able to collect empirical evidence. The SMP member further noted that tiering measures by stage of development might be something that NQF could consider when asking developers to adhere to a certain standard. Dr. Drye noted that this is something NQF has broadly thought about; however, there are some challenges associated with this prospect.

Lastly, Dr. Nerenz noted that the discussions from this meeting tie into the conversations from the April SMP advisory meeting and other conversations from previous advisory meetings about the process and flow of measures through the SMP and beyond in the endorsement process. Dr. Nerenz noted the following as an example: If a measure does not pass the SMP's evaluation due to not meeting a threshold for reliability, it could still move through the endorsement process and be endorsed, but perhaps with a special asterisk noting its importance but also low reliability.

Public Comment

No public or NQF member comments were provided during the measure evaluation meeting.

Next Steps

Gabby Kyle-Lion, NQF analyst, reviewed the next steps. Ms. Kyle-Lion reminded the SMP that NQF staff will create a summary for this meeting. Ms. Kyle-Lion also reviewed the next steps for the reliability thresholds table, which will be to hold a public comment period. The intent to submit deadline for fall 2022 was noted as August 1, 2022. Ms. Kyle-Lion notified the SMP that the NQF SMP team will be sending out a Doodle poll to gauge availability for SMP meetings for the remainder of 2022. Lastly, the list of items for future discussion was reviewed.