Dear EENT Standing Committee,

Since we released the submission materials for measures #2640 and #2811 to you, we have received additional information from the developer that has addressed many of our concerns. Because of the short turn-around time, we are providing this information—as well as our preliminary analysis of the new information—in this "addendum". The developer will officially modify their submission materials after the March 14, 2017 webinar.

--NQF Staff

For both #2640 AND #2811:

- Now specified for only three levels of analysis: individual clinicians ["provider"], clinician practices ["department/group"], and facilities ["institution"]
- ***Still need to clarify*** whether a provider/department/institution must have more than 5 eligible encounters in the measurement time period in order to be eligible for the measure

Measure #2640 [Otitis Media with Effusion - Antibiotics Avoidance]

Reliability

*Updated **Reliability Testing** Results from the developer*

OME, antibiotic avoidance:

Entity	N	F	P
Provider	1,786	26.58	<0.0001
Department/Group*	170	124.9	<0.0001
Institution	6	2,668	<0.0001

* Because of the possibility that providers might rotate among clinics, department/group is conservatively defined as a particular special at a single institution.

NQF Preliminary Analysis

The overall method is appropriate and the updated analysis was conducted for the levels of analysis as specified.

The value 1-1/F can be considered an "average reliability". A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Entity	**N**	**F**	**P**	**1-1/F ("average reliability")**
Provider	1,786	26.58	<0.0001	0.9624
Department/Group*	170	124.9	<0.0001	0.9920
Institution	6	2,668	<0.0001	0.9996

Guidance from the Reliability Algorithm

Specifications are precise (Box 1) → Empirical testing conducted for all three levels of analysis specified (Box 2) → Score-level testing was conducted (Box 4) → Method is appropriate (Box 5) High certainty that the performance measure scores are reliable (Box 5a) → High

The highest possible rating is HIGH.

Preliminary rating for reliability: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Validity

Updated Validity Information from the developer

Data element validity testing: Further information on the analysis of 225 encounters from one site: The data from this site includes 28 primary care practices (4 hospital-based and 24 community-based) with largely practice-specific staff; 21 specialty departments were also included in the OME dataset.

Score-level testing: From the score reliability/discriminant ability testing, we know that sites are different groups from the measure's perspective; this has face validity as well, since we expect that differences in practice and training across institutions will underlie the differences in measure results.

Hypothesis: the same providers will, absent external influences, practice in a consistent way over time.

Rationale for hypothesis:

- The consensus best practice did not change over the interval we're examining (i.e., the specialty society guidelines we're tracking have not changed in respects important to the measure over the interval)
- The technical infrastructure hasn't changed qualitatively (there've certainly been upgrades and optimizations to the EHRs, and to hospital infrastructure, but none at the level of, say, changing EHRs)
- No evidence that there have been attempts to change practice by institutions or payers.

Expected results based on this hypothesis: Measure scores within a given provider (and consequentially for groups of providers) will vary less over time than scores between different providers or groups. In terms of the ANOVA, the hypothesis predicts that the year-over-year F statistic will be smaller than the group-to-group F statistic.

Results of testing:

Entity	F (between entity)	F (between year)
Provider (threshold = 5)	151.3	14.5
Provider (threshold = 10)	21.1	1.9
Department	124.9	1.9

Conclusion: Hypothesis is supported.

NQF Preliminary Analysis

Developers hypothesized that measure results within individual providers/departments would vary less across time than measure results between providers/departments, given the lack of external influences that would affect results across time. This can be considered a form of score-level construct validation.

The analogous analysis for the facility [institution/site] level of analysis was not provided. It is not clear if/how the results of the year-by-site testing analysis that was initially presented support the stated hypothesis. As noted in the initial staff preliminary analysis, that analysis itself would be considered a weak form of construct validation (i.e., comparing the score with itself across time) if the data were not aggregated by provider/department and then by site.

Guidance from the Validity Algorithm

Specifications are consistent with the evidence (Box 1) → Potential threats to validity were assessed (Box 2) → Empirical testing was conducted for at two of the three levels of analysis specified (Box 3) → Score-level testing was conducted for at least two of the three levels of analysis specified [facility-level testing results not clear] (Box 6) → Method is described and seems appropriate (Box 7) → Moderate certainty that measure scores are a valid indicator of quality (Box 8b) → Moderate [ASSUMING site-level results can be explained by the developer]

NOTE: The chart review analysis and the correlation analysis comparing measure results using ICD-9-CM coding versus using a proprietary coding system support the validity of the measure but cannot stand alone because the data were derived from only one EHR (NQF requires testing from more than one EHR).

The highest possible rating is HIGH.

Preliminary rating for validity: ☐ High ☒ Moderate ☐ Low ☐ Insufficient

Measure #2811 Acute Otitis media - Appropriate First-Line Antibiotics

Reliability

*Updated **Reliability** Testing Results from the developer*

AOM, antibiotic appropriateness:

Entity	N	F	P
Provider	2,718	18.21	<0.0001
Department/Group	131	215.6	<0.0001
Institution	6	3811	<0.0001

NQF Preliminary Analysis

The overall method is appropriate and the updated analysis was conducted for the levels of analysis as specified.

The value 1-1/F can be considered an "average reliability". A value of 0 indicates that all variation is due to measurement error and a value of 1 indicates that all variation is due to real differences in provider performance. A value of 0.7 often is regarded as a minimum acceptable reliability value.

Entity	N	F	P	1-1/F ("average reliability")
Provider	2,718	18.21	<0.0001	0.9451
Department/Group*	131	215.6	<0.0001	0.9954
Institution	6	3811	<0.0001	0.9997

Guidance from the Reliability Algorithm

Specifications are precise (Box 1) → Empirical testing conducted for all three levels of analysis specified (Box 2) → Score-level testing was conducted (Box 4) → Method is appropriate (Box 5) High certainty that the performance measure scores are reliable (Box 5a) → High

The highest possible rating is HIGH.

Preliminary rating for reliability: ☒ High ☐ Moderate ☐ Low ☐ Insufficient

Validity

Updated Validity Information from the developer

The developer tested the same hypothesis as noted above for measure #2640.

Results of testing:

Entity	F (between entity)	F (between year)
Provider	18.21	4.7
Department	215.6	1.6

Conclusion: Hypothesis is supported.

NQF Preliminary Analysis

Developers hypothesized that measure results within individual providers/departments would vary less across time than measure results between providers/departments, given the lack of external influences that would affect results across time. This can be considered a form of score-level construct validation.

The analogous analysis for the facility [institution/site] level of analysis was not provided. It is not clear if/how the results of the year-by-site testing analysis that was initially presented support the stated hypothesis. As noted in the initial staff preliminary analysis, that analysis itself would be considered a weak form of construct validation (i.e., comparing the score with itself across time) if the data were not aggregated by provider/department and then by site.

Guidance from the Validity Algorithm

Specifications are consistent with the evidence (Box 1) → Potential threats to validity were assessed (Box 2) → Empirical testing was conducted for at two of the three levels of analysis specified (Box 3) → Score-level testing was conducted for at least two of the three levels of analysis specified [facility-level testing results not clear] (Box 6) → Method is described and seems appropriate (Box 7) → Moderate certainty that measure scores are a valid indicator of quality (Box 8b) → Moderate [ASSUMING site-level results can be explained by the developer]

NOTE: The chart review analysis and the correlation analysis comparing measure results using ICD-9-CM coding versus using a proprietary coding system support the validity of the measure but cannot stand alone because the data were derived from only one EHR (NQF requires testing from more than one EHR).

The highest possible rating is HIGH.

Preliminary rating for validity: ☐ High ☒ Moderate ☐ Low ☐ Insufficient