Looking forward to the Exchange’s future open enrollment periods in years beyond 2013, a set of 74 measures for inclusion in the Quality Rating System (QRS) has been proposed to the Measurement and Reporting Work Group (MRWG). This memo addresses the issue of creating composite scores as part of the proposed QRS by outlining five potential methodologies for constructing composites to summarize individual quality measures, as well as discussing methodological issues related to compositing, such as missing data issues and weighting of measures that comprise the composite.

A composite summarizes two or more individual quality measures into a single quality score. Particularly for public reporting, composites are desirable because they reduce the cognitive task on the consumer to make an assessment of health plan quality based off of numerous individual measures. Implicit in composite design is that the rolled-up individual measures are conceptually similar enough that combining them reflects a broader construct that is an important aspect of health plan quality. The primary downfall of composite scoring is that the resultant score may wash out important quality distinctions at the individual measure level. However, giving consumers the option of drilling down to specific individual measures of interest mitigates this.

In previous memos, the proposed straw model has illustrated how individual measures might be grouped together to form composite scores. Each blue box in the proposed QRS straw model represents a potential composite or sub-composite measure\(^1\). For example, there are individual measures that will roll-up into the sub-composite, *Quality of Customer Service and Claims Processing*, and this sub-composite will then roll into the higher-level composite, *Quality of Plan Features, Services, and Processes*. This memo explores several methodological issues surrounding construction of these composite scores, both for sub-composites and higher-level composites, recognizing that there is no authoritative agreement on the single best method. For this discussion, we assume that the methodology for compositing at lower-levels of the QRS will hold for compositing at higher levels of the QRS as well. Whatever methods are used to create composites should be consistent between all levels of compositing.

This memo will make reference to the “reporting entity” as the unit that will be measured in the QRS, in order to allow for flexibility in terms of the level at which measures will eventually be reported in the

---

\(^1\) See attachment “Proposed Structure for the Minnesota Health Insurance Exchange’s Quality Rating System”
QRS. Based on previous MRWG feedback, the QRS will likely collect data at a product-type level at least in the first few years of public reporting on the Exchange. This means the Exchange will have fewer reporting entities as compared to if data collection occurred at the QHP level. This, in turn, has implications for certain methods of constructing composites as described where relevant below.

Methods for Creating Composites

Method 1: Average Scores

Under this method, measures within a composite domain are averaged to arrive at a reporting entity’s composite score.

Pros: This method is easily understood and results in a numeric score that could potentially be interpretable by consumers – the average percent of times that the appropriate procedures were done for plan members. This method also results in a standard error for composite measures, thus allowing significance testing to be done on the resulting composites.

Cons: This method ignores the differing variation between reporting entities on individual measures. Some measures will have fairly similar scores across reporting entities, whereas others will drastically differ across reporting entities. As a result, measures with larger variances will have a greater impact on the composite score.

Method 2: Average Standardized Scores

Under this method, each measure for each reporting entity is standardized by subtracting the statewide average for the measure and dividing by the statewide standard deviation for the measure. The resulting standardized score represents how many standard deviations the reporting entity is from the statewide average. All resulting standardized measures have the same average (0) and standard deviation (1). The standardized scores for each measure within a composite domain are averaged to arrive at a reporting entity’s composite score.

Pros: Unlike Method 1, all measures have the same impact on the composite score (unless they are weighted otherwise) because the differing levels of variation across measures have been eliminated through standardization of the score. Similar to Method 1, the resulting composite has a standard error and significance testing on the composite is possible.

Cons: The resulting composite score is a measure of how many standard deviations the reporting entity is, on average, from the statewide average. It is not intuitive to consumers. It has a mean of 0 and some reporting entities will have negative composite scores if worse than the average. However, the composite scores could be rescaled to a score more palatable to consumers, such as a 0 to 100 scale, or converted to symbols, as discussed later in this memo under the section, Translating Composite Scores into a Symbolic Rating System.
**Method 3: Converting Scores to Percentiles**

Under this method, each measure for each reporting unit is converted to a percentile based on how it compares to other reporting entities in Minnesota. The percentiles for each measure within a composite domain are averaged to arrive at a reporting entity’s composite score.

Pros: Similar to Method 2, this is a way to standardize scores so that all measures have the same impact on the composite score (unless they are weighted otherwise).

Cons: Converting each measure to a percentile is a way of ranking the reporting entities for each measure. The magnitudes of the differences between reporting entities are lost. This is particularly problematic if there are a small number of reporting entities. This would be an issue for both individual measures and for the composites.

Using the percentile method also ignores the sample size of the reporting entity and the resulting confidence interval around each reporting entity’s score. For example, Plan A may have a higher percentile rank than Plan B, but may not be significantly different than Plan B.

Another disadvantage of this method is that a percentile score has no associated standard error, thus eliminating the option of performing significance tests on the scores.

**Method 4: Converting Scores to Significance Test Results**

Under this method, each measure for each reporting entity is compared to statewide averages and converted to a t-statistic as a test for significance. The t-statistic could then be translated to a score at the individual measure level on a 1 to 3 scale (various confidence levels could be chosen):

- 3=95 percent confident that the reporting entity is above the statewide average
- 2 = not 95 percent confident that the reporting entity is different from the statewide average
- 1= 95 percent confident that the reporting entity is below the statewide average

Or, the score could be translated to a 1 to 5 scale (various confidence levels could be chosen):

- 5=95 percent confident that the reporting entity is above the statewide average
- 4=80 percent confident that the reporting entity is above the statewide average
- 3= not 80 percent confident that the reporting entity is different from the statewide average
- 2 = 80 percent confident that the reporting entity is below the statewide average
- 1= 95 percent confident that the reporting entity is below the statewide average

Other various scales could also be used.

The scaled significance test score for each measure within a composite domain are averaged to arrive at a reporting entity’s composite score.
Pros: Similar to the advantage of Method 2 and 3, this is a way to standardize scores so that all measures have the same impact on the composite score (unless they are weighted otherwise). Furthermore, it helps ensure that the differences in scores reflect actual differences in performance.

Cons: Significance test results depend on both reporting entity scores and sample sizes. Reporting entities with large sample sizes (perhaps due to the fact that they use administrative data) would be more likely to be significantly different from the statewide average as opposed to reporting entities with smaller sample sizes.

The resulting composite score does not result in a standard error and therefore significance testing cannot be performed on the composite score; however, significance testing is in the underlying methodology.

**Method 5: Averaging Symbolic Scores on Individual Measures**

Under this method, each measure would be converted to a symbol based on a combination of the logic used in Methods 4 and 5. As an example, assume a 5-point scale where:

- Reporting entities in the top quartile that were also statistically significantly better than the statewide average would get a 5
- Reporting entities in the top quartile that were not statistically significantly better than the statewide average would get a 4
- Reporting entities which are in neither the top nor bottom quartile would get a 3
- Reporting entities in the bottom quartile that were not statistically significantly worse than the statewide average would get a 2
- Reporting entities in the bottom quartile that were statistically significantly worse than the statewide average would get a 1

The assigned score for each measure within a composite domain are averaged to arrive at a reporting entity’s composite score.

Pros: This method takes into account both the statistical significance of a reporting entity’s score compared to the statewide average as well as the substantive difference of a reporting entity’s score compared to the statewide average.

Cons: The magnitude of the differences between reporting entities is lost when translated into quartiles. This is particularly problematic if there are a small number of reporting entities. This would be an issue for both individual measures and for the composite measures. In addition, the composite score itself has no statistical properties for significance testing.

**Weighting Individual Measures and Sub-composites**

Weights allow different measures to be given greater or lesser importance in their contribution to a composite score. Decisions about weighting must be made both in how individual measures are
weighted in a sub-composite category as well as in how those sub-composites are weighted in an aggregate composite. Giving equal weight to all measures, while common, may not be the best approach to calculating a composite if measures vary in their ability to assess health plan quality and in their salience to consumers. These factors often cannot be quantified based on the data provided by plans, but can be estimated based on:

- A measure’s eligible denominator size
  - Some measures address aspects of quality that affect more plan members than others. For example, under the Adults Staying Healthy composite domain, more enrollees will be eligible for inclusion in the Flu Shot denominator than will be eligible for inclusion in the Advising Smokers to Quit denominator.

- Expert assessment of the measure
  - Measures were proposed for inclusion in the QRS having considered criteria such as importance to, and potential to improve, health care quality. Measures that have been endorsed by NQF have been proven to meet these criteria and thus NQF endorsement was weighed heavily in the decision to even propose a measure for inclusion in the QRS. The level to which each measure falls within these criteria can also be used to assess how much weight the measure should be given within the QRS. Expert analysis can be used to rank measures and consequently determine each measure’s weight by assessing:
    - How significant the effects of the measure are on factors such as morbidity, mortality, general quality of life, and satisfaction with health care;
    - How much room for improvement the measure possesses; and
    - How strongly measure improvement is related to changes in overall health care quality.

- Assessment of the sub-composite or measure based on general consumer interest
  - Consumers will have differing opinions on which sub-composites are important to them. Proposed sub-composites will be evaluated as part of the Exchange’s consumer testing process. The Work Group will have an opportunity to hear the results of consumer testing and consider adjusting proposed weighting in response to consumer interest.
Handling Missing Data

There may be some measures in the QRS that will be missing for a reporting entity due to lack of data. This would most likely be due to the reporting entity not meeting the minimum sample size to be considered for public reporting; Missing data will affect the way measures roll up into composites and there are various ways of handling missing measures within a composite.

Option One: Omit the missing measure to calculate the composite score.

• If a reporting entity is missing a measure from the composite domain, then drop the measure from the calculation of that reporting entity’s composite score. For example, if there are three measures in a composite domain and one is missing, average the results of the remaining two measures to calculate the composite score.
• This option can be easily incorporated for all proposed compositing methods, with the exception of Method 1: Average Scores. Since the scores are not standardized, omitting a measure for some reporting entities and not others would result in a comparison based on differing methodologies.

Option Two: Assign the average score to missing measures to calculate the composite score.

• For a given missing measure, assign the reporting entity a score of the measure’s statewide average. Then use this assigned score when averaging the measure results within the composite domain to calculate the composite score.
• This option could be used for any of the proposed compositing methods.
• This option will heighten a typically low-performing reporting entity’s score, while lowering a typically high-performing reporting entity’s score.

Multiple options may be used depending on the reason for missing data. Whatever method (or combination of methods) is used, we recommend a minimum number of individual measures within the composite domain that a reporting entity must reach for the composite score to be calculated. This ensures that a reporting entity’s composite is not misrepresented by the presence of too few individual measures for the reporting entity. For example, a reporting entity may need to be reportable on at least 50% of the measures for a composite score to be calculated. If a reporting entity does not meet this criterion, then an explanatory statement such as “Not enough data to report” may be shown depending on the circumstances surrounding the composite’s missing data. The threshold for minimum number of
measures required should be determined after measure weights within composites have been established.

**Calculating Higher-Level Composite Measures**

Sub-composites can be rolled-up into higher-level composite measures including an overall quality of health plan composite. Sub-composites can be combined into a higher-level composite in the same way that measures were combined into sub-composites with the following limitation - if the method for combining the sub-composites requires a statistical test of significance, a method of combining measures into sub-composites that yields a standard error for the sub-composite must be used.

Individual measures could also be combined directly to arrive at a higher-level composite – including an overall quality of health plan composite. Using any of the five methodologies discussed, combining individual measures to arrive directly at a higher-level composite will return the same results as combining sub-composites into a high-level composite in the absence of missing data.

If there is missing data, combining measures directly into a higher-level composite may enable the higher-level composite to retain information from an individual measure that would be lost if sub-composites were combined into higher-level composites. This is because, depending on the decision rules implemented regarding how much missing data will be allowed, a sub-composite for which some but not a sufficient number of measures are available will not be calculable.

**Translating Composite Scores into a Symbolic Rating System**

Composite scores are commonly presented as symbols to consumers for ease of interpretation. These symbols could be stars, upward and downward point arrows, or other easily understandable symbols of quality. A composite score can be converted to symbols in a number of ways, depending on the methodology used to calculate the composite score.

*Option One: Rankings/Percentiles/Quantiles*

- Composite scores are ranked and cut points are drawn based on statewide results. For example, the top third performing reporting entities get 3 stars, the second third gets 2 stars, and the bottom third gets 1 star.
This option of presenting composites forces a predetermined distribution of reporting entity scores (e.g. there must be reporting entities that get one star and there must be reporting entities that get 3 stars).

- Differences in stars reflect a rank order difference but not necessarily a statistically meaningful difference.
- This option can used under **Method 1: Average Scores, Method 2: Average Standardized Scores, or Method 3: Converting Scores to Percentiles**.

**Option Two: Significance Tests**

- Composite scores are compared to a statewide average and scores that are statistically significantly higher than the statewide average are given the symbol for “above”; scores that are statistically significantly lower than the statewide average are given the symbol for “below”; scores that are not significantly different from the statewide average are given the symbol for “average.”
  - Under this option, both the composite score and standard error impact the final display.
  - The data dictates the number of reporting entities above and below average.
  - Reporting entities with larger sample sizes are more likely to have scores reported as “above” or “below.”
  - This option can be used under **Method 1: Average Scores or Method 2: Average Standardized Scores**.

**Option Three: Combination of Quantiles and Significance Tests**

- Both a reporting entity’s ranking and its statistical significance play a role in identifying the reporting entity’s final symbolic score. For example, reporting entities in the top 25% that are statistically significantly above the statewide average would get 5 stars; reporting entities in the top 25% that are not statistically significantly above the statewide average would get 4 stars; reporting entities in the bottom 25% but are not statistically below average would get 2 stars; reporting entities that are in the bottom 25% that are statistically significantly different from the statewide average would get 1 star; and reporting entities that are neither in the top or bottom 25% would get 3 stars.
• This option could be approached in a different way where reporting entities that are statistically significantly different would be the primary driver of the number of stars and the reporting entity’s ranking would be the secondary determination of stars.
• This option can be used under Method 1: Average Scores or Method 2: Average Standardized Scores.

Option Four: Averaging Measure Level Symbols

• If individual measures are assigned a symbolic score, as in Method 4: Converting Scores to Significance Test Results and Method 5: Averaging Symbolic Scores on Individual Measures, averaging those scores could be presented as a composite measure. For example, if reporting entities are assigned 1 to 5 stars for each measure, the composite score could be presented as the average number of stars within the composite domain.
• This option would lend itself to non-whole number composite scores (e.g. 3.7 stars).

Challenges Associated with Compositing of Certain Proposed QRS Measures

Some measures in the proposed QRS lend themselves more readily to certain methods of roll-up to a composite score. For example, most of the proposed HEDIS/CAHPS measures are proportions of how often quality performance occurred. This kind of measure has an associated sample size, and consequently a variance around the resulting proportion. This characteristic of variance allows the score to be used in significance testing against the statewide average, or against other reporting entities. Any of the above methods for calculating composites can be applied to these measures.

Another subset of the proposed QRS draws on the eValue8 measures. These measures are also proportions, but the proportion is a percentage of total points available for the measure. For example, under the Prevention & Health Promotion measure, there are 168 earned points possible using the full 2012 eValue8 scoring. A plan can receive points by having programs and processes in the areas of Activities to Support Healthier Birth Outcomes, Prevention and Treatment of Obesity, Prevention and Treatment of Tobacco Use, Immunization Programs, Cancer Screening Programs and Results, and Worksite Health Promotion and Health Assessments². Since the proportion is a reflection of total points earned rather than a reflection of how often quality occurred (as in the HEDIS/CAHPS measures), there is no concept of sample size or variance around the eValue8 measures. The RAND Cultural Competency

---

Implementation Measure is similar to the eValue8 measures in this regard. Thus, the proposed composite methodologies that rely on significance testing would not be applicable for these measures. However, these measures could be composited using Method 1: Average Scores, Method 3: Converting Scores to Percentiles, or Method 5: Averaging Symbolic Scores on Individual Measures.

There are a few proposed measures that are proposed for inclusion in the QRS because they addressed desired aspects of plan performance that were not addressed by the HEDIS/CAHPS or eValue8 measures. These aspects of quality include resource utilization as an indicator of cost, as well as network adequacy. Although health plan measures exist that address these aspects of performance, they may not fit well into composites of a QRS.

For example, the relative resource use measures show how much care of various types is used in a plan for members with certain conditions relative to the amount that would be expected to be used after adjustment for member characteristics. The resulting ratios of actual to expected can be reported for each condition (diabetes, heart disease, etc.) and compared from plan to plan. The utilization measures also summarize the use of services by plan members for different types of care (ambulatory services and inpatient services). These measures are mathematically different than the proportions that come from HEDIS/CAHPS-type measures. More important, the relevance to consumers will generally be different from the relevance of the other quality measures. In fact, the underlying concept of resource utilization measures is that they will be used alongside quality measures so that users can assess whether some plans are achieving better quality with lower cost and resource utilization. The resource use measures are of particular relevance for plan management and to help purchasers evaluate and influence plan management. From a consumer's standpoint, low resource use might have implications for likely out-of-pocket costs, but that type of interpretation is more likely to be relevant in the context of a comparison of premium and out-of-pocket costs than in the construction of composite measures of quality.

Thus, the resource use measures as they stand by themselves do not necessarily reflect high or low plan quality. Rather, these measures are useful when compared to health plan quality so as to give the consumer a better idea of value (e.g. high resource use and low quality would mean poor health plan value, whereas low resource use and high quality would mean good health plan value). It would not be recommended to combine a measure of resource use into composites with other proposed measures of health care quality.
Network adequacy was also identified as a gap in desired aspects of health plan quality, and two URAC measures were proposed to address this gap. These measures are derived from CMS Medicare Part C reporting requirements, and assess the number of specialists and primary care physicians accepting new patients. Currently, these measures are not being used in a public reporting format and the measures themselves are not proportions, but raw numbers of providers. Although this information may be useful to consumers in deciding how much flexibility they may have when choosing a provider with a plan, given how different the format of these measures is from the rest of the proposed measures, it would not be recommended to roll these measures into a composite score. Please note these measures are still being considered for inclusion in the Quality Rating System and may be reported separately from the composite quality rating indicator.

Discussion Questions:

• What method should be used for creating a composite measure?

• What approach should the Exchange use in determining weights for individual measures and/or for subcomposites?

• How should the Quality Rating System address issues associated with missing data?

• What method of comparing reporting entities should be used?