



## PUBLIC HEALTH ACCREDITATION BOARD

### SCORING AND WEIGHTING

January 15, 2009

Released by the Board of PHAB for Vetting

#### OBJECTIVES TO BE ACHIEVED BY THE SCORING AND WEIGHTING METHODOLOGY

- **Construct/Content Validity:** Ensure that the response options chosen for use with each performance measure provide an accurate representation of the activity being measured. Also ensure that the methodology chosen for weighting and constructing composite measures preserves this validity.
- **Discriminant Validity (differences across agencies):** ensure that the response options chosen for use with each performance measure allow detection of meaningful differences in performance across agencies. Also ensure that the methodology chosen for weighting and constructing composite measures preserves this type of discriminatory ability.
- **Discriminant Validity (improvement over time):** to the extent possible, ensure that the response options chosen for use with each performance measure allow detection of meaningful improvements (or decrements) in agency performance over time. Also ensure that the methodology chosen for weighting and constructing composite measures preserves this type of discriminatory ability. It is important to note that a measure's ability to detect differences in performance across agencies does not necessarily correspond to its ability to detect improvements in performance over time.
- **Reliability:** ensure that the response options chosen for use with each performance measure can be applied consistently by different reviewers (inter-rater reliability) and can be applied consistently from one time period to another (repeatability).
- **Feasibility:** ensure that the response options chosen for use with each performance measure are feasible to implement based on the volume and nature of evidence that an accreditation reviewer must evaluate in order to assign the appropriate response option.

## PROPOSED SCORING AND WEIGHTING

### **(1) SCALE: Both a five point scale and a smaller, three-point sub-scale will be used and compared during Beta testing.**

#### **Considerations:**

- A larger number of response options (e.g. five-point scale) potentially allows for greater **discriminatory power** in detecting differences in performance and possibly for detecting improvements in performance over time. Ceiling effects and/or floor effects may be diminished with multiple response options.
- However, too many response options for a given measure may generate false precision (i.e. create the appearance of differences in performance where there are none in reality) and result in **low reliability** by causing reviewers to face uncertainty about which response option best fits a particular agency.
- The Standards Committee recommended using a three point scale to strike a balance between reliability and discriminant validity. A three point scale might put many responses in the middle. However, a three point scale would be feasible as an initial scoring strategy, allowing for some partial documentation.
- The PHAB Board considered the use of a five point scale so that two additional scoring levels are available to recognize those agencies that go above and beyond an “accredited” level of performance by actively engaging in QI for a given measure. The Beta test can be used to compare the performance of this five-point scale with a smaller three-point scale that collapses the top 3 categories into a single category.

#### **(2) RESPONSE Considerations:**

- Item response theory suggests that response categories—and the boundaries between these categories—should be defined so as to emphasize meaningful and observable differences in performance or level of achievement, as indicated by available response data. Please note that if a 5-point scale is selected for accreditation, the options below will likely be expanded. Options may include:
  - **[No Activity, Partial Compliance, Full Compliance]**: in this option, the bottom response category is reserved for agencies that do not have any activity to report in a given performance area. Another way of wording these response options could be: [Meets None of the Criteria for this Measure, Meets Some But Not All of the Criteria, Meets All of the Criteria for this Measure]
  - **[Low Performance, Moderate Performance, High Performance]**: in this option, the bottom category includes agencies with “no activity” as well as agencies with limited or “low” activity levels in a given performance area. The distinction between the bottom category and middle category may be more difficult to operationalize under this option than under the first option.
- The first option provides a relatively clean and observable distinction between no activity, partial achievement, and full achievement. The distinction between the top two categories could be operationalized for each measure by specifying the types of evidence needed to justify a “partial” versus “full” rating.
- Washington State started with using the word “compliance but changed to “demonstration.” The Washington language is currently “demonstrated the elements of the measure, partially demonstrated, or not demonstrated.” This has a positive connotation, in

that the department presents their work to demonstrate their meeting the measure and the department is being judged on what is presented. “Compliance” implies that the standards and measures are absolute. “Performance” is too vague a term.

**(3) MINIMUM THRESHOLD: Accreditation should be conferred only if an agency meets a minimum threshold of performance in all domains.**

**Considerations:**

- The standards development workgroup recommended that minimum thresholds should be established at the domain level, but there should be no specification of which individual measures within a domain must be met. Specifying individual measures that must be met could be problematic for domains with relatively small numbers of measures and for domains with very different numbers of measures.

**(4) OTHER ISSUES: The beta test will test:**

- a. Alternative methods of assigning numeric values to response categories to assess the sensitivity of results.**
- b. The implications of weighting schemes by assessing the sensitivity of results to alternative weighting scenarios.**
- c. Numeric thresholds to use for a minimum score, by testing a variety of scenarios with data generated through the beta testing process.**
- d. Both a three point scale and a five point scale.**