Appendix C –Approach to Observational Measurement Development

Evaluating reliability and validity of data from observation-based measures of settings requires cautious application of standard psychometric concepts and tools (Cronbach et al, 1963; Raudenbush and Sampson, 1999; Seidman, in press) and careful alignment between (a) the different purposes for which scores will be used and (b) the different methods to determine score reliability and validity.

Specific challenges include the following.
The instructional practices recommended by experts may not occur in all settings all the time. Observational measures and methods of data collection that are not calibrated to offering structure and sequence may both miss critical practices that do in fact occur or, produce low scores for practices which are not part of the curriculum.
Many setting-level measurement constructs are formative rather than reflective in nature, meaning that the items grouped within a given scale may not "reflect" a construct that exists independently of the items. Formative constructs do not necessarily exhibit "internal consistency" among items and are better understood as indexes.
Facets of data collection – items, raters, time of day and year, programs, and interactions of these facets, may introduce substantial error into quality scores. These sources of unreliability can only be detected with data collection designs that "cross" raters.
There is often pressure to improve score reliability, even when at cross-purposes with more important goals for validity. For example, a single total score with high internal consistency, high construct validity, and low rater bias may be achieved by deleting many items from the Youth PQA and may serve purposes of differentiating between high and low quality sites. However, for learning and behavior change purposes less reliable scores that describe specific staff behaviors or sets of practices that typically co-occur may be more useful.

For these reasons our approach to the development of observational measures consists of the following steps:

Step 1. Content and Substantive Validity – Which instructional practices are important and where can an observer see them? Both measures and data collection methods can be adjusted to maximize opportunities to observe instructional practices of specific interest. This step involves literature review, consultation with expert practitioners, drafting items, empirical analyses to see how items group, and asking practitioners when and where we may see these practices.

Step 2. Reliability – Do multiple raters produce the same score? Our goal in this step is to maximize inter-rater reliability at the item level. Our primary analytic tools include qualitative analysis of rater reflections on the meaning of language in items, percent perfect agreement, and intraclass correlation coefficients (ICC). Internal consistency as a measure of reliability for multi-item scales is only appropriate for reflective scales. In a reflective scale, each item is theorized to "reflect" a latent construct – with interchangeability of items assumed – and any item should provide a reflection of the underlying construct; the latent construct is assumed to "cause" the item responses. For observation-based measures of behavior, however, groupings of items are most often formative in the sense that the items add up or "form" the composite score.[9] Scores for formative measures are best constructed as sum scores or indexes, and are best evaluated by reference to inter-rater reliability (measures of internal consistency are not appropriately applied).

Step 3. Convergent Validity — Are observation-based scores associated with other relevant measures? Convergent validation demonstrates how quality scores relate to other measures implicated by our theories of organization and child-level change (See Appendix D). Because relationships between fine-grained measures of teacher behavior (e.g., planning or reflection) are (a) not specified clearly by research and (b) likely to be context dependent[10], we are frequently interested in point-in-time relationships between a total score (e.g., the setting features many good staff practices) and other policy and theory relevant constructs such as teacher education and youth engagement. Guided by theory, we employ both linear and pattern-centered analytic methods to investigate point-in-time patterns of association.

Step 4. Contribution of Methods to Unreliability - How do facets of data collection method produce measurement error? Following steps 1-3, we use techniques drawn from generalizability theory to understand systematic error associated with several facets of data collection method including items, raters, time of day and year, program type, and interactions among facets. Analysis of variance methods estimate true score and error based on data collection designs that "cross" the several facets of method and use methods that maximize score reliability (e.g., more raters, more days).

Although only summarizing our approach to reliability and validity, these steps support recommendations for use of observation-based measures in lower stakes circumstances for performance feedback and continuous improvement.

[9] For more information, see Bollen (1984); Diamantopoulos and Siguaw (2006); Jarvis, MacKenzie, and Podsakoff (2003)..
[10] For more information on interpretational confounding, see Hardin, Chang, Fuller and Torkzadeh, 2011.