Development of the Academic Skill-Building Program Quality Assessment

Summary

The *Academic Skill Building Program Quality Assessment* (ASB PQA) was designed to serve as a performance standard and quality measure for tutoring and other academically focused afterschool programs. As such the ASB PQA is part of a broader suite of Program Quality Assessments that align to the *Youth Program Quality Intervention* and 105 quality improvement systems including over 4,000 out-of-school time programs in 38 states. The ASB PQA is an observation-based measure of staff practices related to academic learning and explicit skill development objectives that entail disciplined engagement with content. The ASB PQA Measure Development Study entailed development and pilot testing of 22 new items focused in domains of instructional practice including: (1) targeting student focus on specific academic content and supporting student practice; (2) sequencing of student experience to assure exposure to incremental task design, models, and guided instruction; (3) application of learning strategies and effort-achievement mindsets; (4) higher order reflection and evaluation of content and skill development. The study used a mixed methods design to engage expert practitioners in item development, develop rater training, conduct data collection for 12 tutors in four programs, and gather feedback from a more diverse sample of 23 afterschool providers. Results describe aspects of reliability and validity for ASB PQA data and provide recommendations for next steps in the development of the measure.

Findings and Outcomes

This measure development study was designed to cost-effectively advance our understanding of best practices for tutoring and other academically focused out-of-school time settings – and to increase capacity to improve them by making the ASB PQA available for use in continuous improvement approaches.

Project outcomes included development of 22 new ASB PQA items and rubrics in collaboration with 31 expert practitioners who provided detailed feedback throughout the iterative process of literature review, item development, data collection, and revision. Another important project outcome was the development of supports (training and guidebook materials) to improve reliability of ratings produced by both external raters and staff teams completing program self-assessment. While the design, iteration, and validation sequence conducted through this study is only a first step, we believe that the ASB PQA (David P. Weikart Center for Youth Program Quality, 2014), the ASB Handbook (Smith, Hillaker, & McGovern, 2014), and the final report for the project (Smith, et al., 2015) provides sufficient information for effective use of the ASB PQA in lower stakes quality improvement systems. This study provides a foundation for

further improvements to the ASB PQA measure and implementation supports as time and resources allow.

In addition to the measure and supports, a final outcome from the study was a set of findings regarding the reliability and validity of the ASB PQA measures as well as a first empirical description of tutoring settings using the ASB PQA. The sample for these analyses included four tutoring programs and12 unique tutors receiving a paired rating on two occasions over a two week period for a total of 48 observations. This intensive sample allowed us to minimize cost associated with data collection while setting baselines for reliability which we will continue to improve over time.

Evaluating reliability and validity of data from observation-based measures of settings requires cautious application of standard psychometric concepts and tools (Cronbach, Nageswari, & Gleser, 1963; Raudenbush & 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 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 often better understood as indexes of items that demonstrate reliability through inter-rater agreement and test-retest agreement at the item level.
- Facets of data collection – items, raters, time of day and year, programs, and interactions of these facets - may introduce substantial error into quality ratings. These sources of unreliability can only be detected with complex data collection designs that "cross" raters and other important facets of observational measurement so that score variance may be partitioned.

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 ASB PQA and may serve purposes of differentiating between high and low quality sites. However, for learning and behavior change purposes scores that describe specific staff behaviors or sets of practices that typically co-occur may be

more useful. This is particularly the case where performance ratings for staff practice are primarily used as feedback to the program site where the data was collected (N=1) and where the local context is known by the interpreters of the data and used to inform the meaning of the performance scores. Prior evidence from the Youth Program Quality Intervention Study suggests that PQA-type measures can be used as effective performance supports in a lower-stakes quality improvement system, despite challenges of measuring reliability for specific items.

For these reasons our approach to the assessment of the reliability and validity of ASB PQA consisted of a set of steps, following the Weikart Center's approach to the development of observational measures (Smith, Hallman, et al., 2012). These steps were designed to maximize our understanding of these complex issues within the limitations imposed by the project budget. Specifically, we first examine the content validity of the ASB PQA by engaging experts around the importance of instructional practices named in ASB PQA items. Next we collect data and examine various aspects of reliability for items, scales and total score. Finally, we draw upon other data sources to examine various patterns of convergent validity.

Findings from the study included the following:

1) *The ASB PQA method was successfully implemented during tutoring sessions*. Four trained raters produced 52 sets of quality scores during 28 unique tutoring sessions. Each quality score was produced following one hour of observation and experience raters required under an hour to score the measure.
2) *The ASB PQA items were characterized as having high content validity by expert practitioners from tutoring programs and from academically oriented afterschool programs*. We asked two groups of expert raters from structured tutoring programs (N=8) and a diverse set of afterschool programs with academically focused offerings (N=23) to evaluate each item in terms of importance to high quality tutoring and the prevalence with which the item is likely to be observed during a typical session. The final set of 22 ASB PQA items were rated as both important and, in most cases, a regular part of good tutoring practice.
3) *A majority of ASB PQA items demonstrated acceptable levels of reliability*. Fourteen of twenty-two items were both of high importance to experts and exceeded the acceptable level on at least two of the three reliability indicators described below. Two more items were both rated high importance and exceeded acceptable reliability on one of the three indicators. All items on the

Learning Strategies Scale had low reliability coefficients. Specific item-level reliability findings include:

a. Exact inter-rater reliability defined as exact agreement between paired raters was in the acceptable range (>80% perfect agreement) for 10 of 22 items. Eight items fell below 70% perfect agreement.
b. Relative inter-rater reliability defined as the ratio of score variance within rater pairs to total score variance was calculated using an intra-class correlation. Ten of twenty-two items were in the acceptable range (ICCs $\geq$ 0.6). Six items were below the minimum interpretable ICC value (ICC $\leq$ 0.4).
c. Test-retest reliability was assessed using a correlation coefficient between time 1 and time 2 (two weeks apart) scores for each of 12 tutors. Nine of 22 items were in the acceptable range ($r \geq 0.7$).

Sixteen of the twenty-two ASB PQA items are organized as formative scales or indexes. We also constructed an ASB PQA total score as the evenly weighted average across all twenty-two items. Scale-level findings include:

4) *Most ASB PQA scales demonstrated acceptable levels of reliability*. Three of the four scales behave more like reflective scales in producing near-acceptable levels of internal consistency, relative inter-rater reliability, and test-retest stability coefficients – and it is likely that these coefficients could be improved with elimination of a few very poorly performing items. The Learning Strategies scale produced very low reliability coefficients on all counts and is clearly not a reflective scale.
    a. Despite the fact that the ASB PQA scales are formative in nature, we calculated Cronbach's alpha for each of the scales and the total score: Targeted Learning (4 items, $a = 0.57$), Scaffolding (4 items, $a = 0.69$), Learning Strategies (3 items, $a = 0.43$), Higher Order Thinking (5 items, $a = 0.65$), Total Score ($a = 0.88$).
    b. Inter-rater reliability defined as the ratio of score variance within rater pairs to total score variance was calculated using an intra-class correlation. Three of four scales and the Total Score were in the acceptable range (ICCs $\geq$ 0.6).
    c. Test-retest reliability was assessed using a correlation coefficient between time 1 and time 2 (two weeks apart) scores for each of 12 tutors. Three of four scales and the total score were in the acceptable range (ICCs $\geq$ 0.6).

5) *The ASB PQA total score demonstrated moderate evidence of convergent validity*. The ASB PQA total score had a moderate positive correlation with the Youth PQA Instructional Total Score and the level of expertise of the tutor.
6) *The quality profile of tutoring programs differs from enrichment oriented afterschool programs*. Following our theory that tutoring programs have different program designs than afterschool enrichment programming, comparison of quality profiles for tutoring versus academic enrichment offerings reflects support for this theory. Enrichment programming is more focused on active learning, qualities of youth-youth interaction, and choice while tutoring environments are lower on each of the quality domains.

Detailed discussion of outcomes and findings from the ASB PQA Measure Development Study can be found in Smith, Hillaker, Thompson, Gersh, and McGovern (2015).

Relevance

The sequence of design, iteration, and validation toward development of effective interventions and measures is long and expensive. At the outset, it is critical to note that our focus in this small project was on content validity – naming the important parts of high quality tutoring practice – and on development of infrastructure for use of the performance measure. In short, at this early stage of the measure development process we are willing to accept lower levels of measurement precision while trying to identify a full range of best practices for tutoring instruction and building out infrastructure to support use of the measure in the field. While this measure development study was a first step in the development of the ASB PQA, there is clearly a hunger for formative assessments like the ASB PQA in the broader out-of-school time field.

However, significant challenges exist in the development of performance measures with acceptable levels of reliability and validity. Frequently, researchers are discouraged from moving forward with observation-based performance measures because traditional psychometric approaches to measure evaluation produce poor results. We would also suggest that these approaches may at times be inappropriately applied. The Weikart Center's approach to development of observational measures may be of interest to broader discussions of methodology for performance improvement.