



# Development and Early Validation Evidence for an Observational Measure of High Quality Instructional Practice for Science, Technology, Engineering and Mathematics in Out-of-School Time Settings

The STEM Supplement to the Youth Program Quality  
Assessment

*Report to Providence Afterschool Alliance*



---

# Development and Early Validation Evidence for an Observational Measure of High Quality Instructional Practice for Science, Technology, Engineering and Mathematics in Out-of-School Time Settings: The STEM Supplement to the Youth Program Quality Assessment

*Prepared by*

Charles Smith  
Samantha Hallman  
Barb Hillaker  
Samantha Sugar  
Gina McGovern

The David P. Weikart Center for Youth Program Quality, A Division of the Forum for Youth Investment



DAVID P. WEIKART  
CENTER FOR YOUTH  
PROGRAM QUALITY

the  
forum  
FOR YOUTH INVESTMENT

*In collaboration with*

Elizabeth Devaney  
Providence After School Alliance



## Introduction

During the summer of 2011, the Providence After School Alliance (PASA) sponsored The AfterZone Summer Scholars Program – a program in which nine community-based organizations partnered with Providence School District teachers and PASA to develop and deliver nine different summer courses from the fields of environmental science, technology, engineering, and mathematics (“STEM”) to 250 Providence Middle School youth. Ten student cohorts, each consisting of 25 students, spent two days per week on site at a school focusing on intense math instruction and two days per week in the field with one of the nine partner organizations. Two instructors moved with them at all times from the classroom to the field and back to make connections between the two settings.

In order to evaluate the AfterZone Summer Scholars model and collect information for future improvement, PASA (a) hired an external evaluator for the project; (b) committed to providing continuous improvement supports to participating program managers and content providers (quality assessment and coaching); and (c) formed an evaluation advisory board to monitor the development and implementation of the external evaluation. In addition, PASA contracted with the David P. Weikart Center for Youth Program Quality (Weikart Center) at the Forum for Youth Investment to develop an observation-based measure of instructional practices to support continuous improvement during STEM programming. This report describes the process of development of the STEM supplement to the Youth Program Quality Assessment (Youth PQA; HighScope, 2005) and preliminary reliability and validity evidence based on data collected during AfterZone Summer Scholars program.

## AfterZone Summer Scholars Program Overview

### Program Description

The 2011 AfterZone Summer Scholars Program, an extension of the school-year AfterZone program that serves middle school students across Providence, was a four-week program integrating hands-on, project-based STEM learning activities. The program ran Monday through Thursday from 8:30am until 1:00pm and included transportation and meals. Youth could enroll in one of ten possible cohorts or “courses”, each of which was led by a team of educators including a middle school teacher, a community educator from a STEM-related organization, and an AfterZone youth worker. Students spent two days in the classroom at a school site receiving instruction from math and English language arts teachers and two days in field settings engaged in hands-on, inquiry based STEM activities applying their classroom learning. The middle school teacher and AfterZone youth worker were consistent instructors, traveling with them from the classroom to the field locations and back to the classroom to help draw connections across settings. Course names and descriptions appear in Table 1.

**Table 1. AfterZone Summer Scholars Program Offering Descriptions**

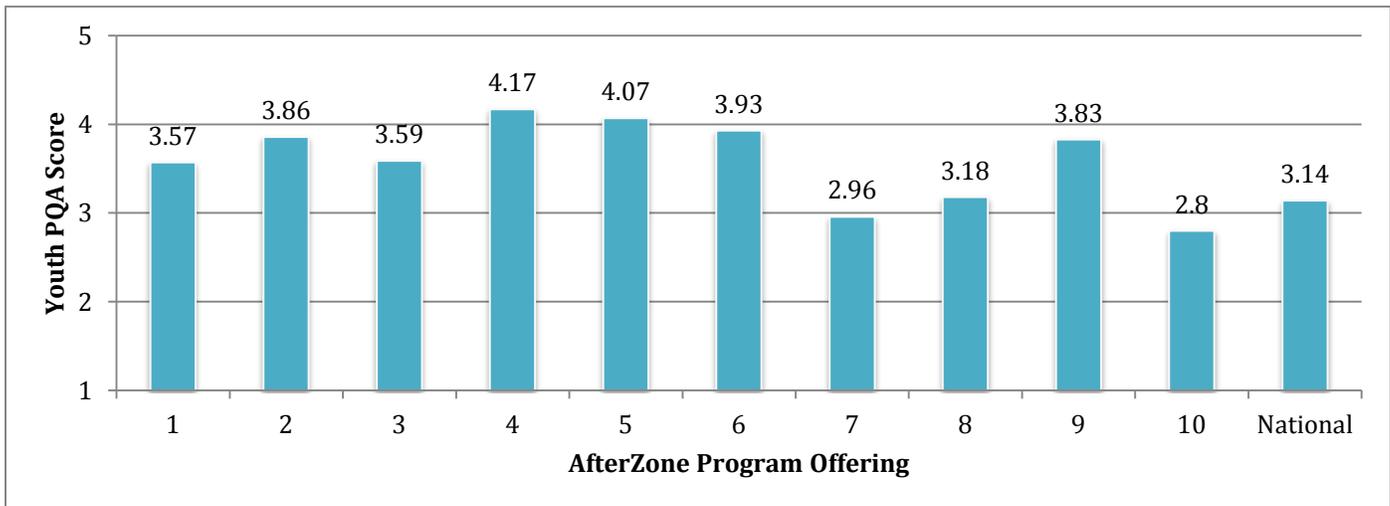
Name	Description
Green Up Summer	Lessons in renewable energy, garbage & recycling, and visits to facilities that process and distribute food and water for consumption
Museum of Natural History Summer Science Fun	Hands-on activities and engineering challenges that encourage investigation in the forces that shape the earth
DownCity Design: Build It	Lessons in design and construction followed by the planning and building of a permanent structure for Providence Park
Explore the Bay 1	Lessons in marine science followed by boat trips in Narragansett Bay
Explore the Bay 2	Lessons in marine science followed by boat trips in Narragansett Bay
Art Explosion	Lessons in how artists use math and science to create works of art followed by creation of sculptures, paintings and tile mosaics
River Adventures	Lessons in watershed ecosystems, experiments in water testing and canoe adventures
Sailzone	Lessons in how to sail a boat followed by discussion of the math and science involved in sailing
Mechatronics	Design and construct electric go-carts using engineering machinery and welding tools
Urban Naturalist	Lessons in urban ecology and exploration of animals and plants in the neighborhood

## Program Quality

The quality of instruction and other setting characteristics during STEM offerings was assessed using the Youth Program Quality Assessment (Youth PQA; HighScope, 2005) – a standardized, observation-based measure that provides a score (1, 3, or 5) based on the degree to which a practice/characteristic was present for all youth in the setting (see Appendix A for Youth PQA content). Reliability and validity of this instrument have been evaluated in several studies (Blazevski & Smith, 2007; Smith & Hohmann, 2005; Smith et al., 2012; Smith, Peck, Denault, Blazevski, & Akiva, 2010). The items included in the Youth PQA overlap substantially with the AfterZone’s identified

“Critical Elements of Inquiry-Based Learning in Informal STEM Education Settings” (see Appendix B). The use of the Youth PQA therefore provides alignment between the definition and measurement of “high quality” inquiry based instruction in AfterZone settings. Two paired observations were conducted in each of the 10 programs for a total of 40 observations by trained raters (data collection described in a later section). Overall, AfterZone program offerings exhibited high quality (higher scores indicate higher quality offerings), with eight of the ten programs scoring higher than normative scores in the Weikart Center’s non-representative national database<sup>1</sup> (see Figure 1).

Figure 1. AfterZone Summer Scholars Program Offering Youth PQA Scores



## STEM PQA Development and Pilot

### Development of the Draft Measure

The Weikart Center employs an intentional approach to observational measurement construction. This approach, detailed in Appendix C, is designed to address numerous challenges associated with using observation-based rubrics to quantify instructional behavior and other setting characteristics. The approach fits within our broader theoretical perspective on the way that out-of-school time settings influence youth development, which is described in Appendix D.

Three steps guided the content and structure of the STEM supplement to the Youth PQA. The first step was to address guidelines developed by PASA’s STEM Community of

Practice<sup>2</sup> for “inquiry based learning” (see Appendix B). Upon review of these guidelines, Weikart Center staff concluded that they aligned substantially with the original Youth PQA and, as such, additional STEM measures should focus on instructional practices more specific to the four STEM disciplines – science, technology, engineering and math – as delivered in out-of-school time settings.

In an effort to define these more specific practices, a literature review was conducted to identify instructional practices associated with development of both content and procedural knowledge, and instructional practices associated with the facilitation of student interest and efficacy in the STEM field (see Appendix E for literature review). Using the literature as a guide to instrument content and structure, several items thought to capture program quality in a STEM context were created, and then organized into scales based on their theorized relationship with each other<sup>3</sup>.

<sup>1</sup> The full reference sample includes 902 offerings collected in programs in nine states. This sample was drawn from a universe of organizations that delivered year-round programming, had full-time administrators, and could produce a weekly schedule of offerings. Overall, 35% of the offerings served primarily elementary school youth, 42% served primarily middle school age youth, and 23% served high school age youth. Thirty-eight percent of offerings were located in community based organizations, 56% were located in school based programs and 7% were in camps.

<sup>2</sup> This group is made up of STEM providers (those involved in the Summer Scholars program as well as others), Rhode Island College faculty, and middle school teachers from the Providence District that has met regularly since 2010.

<sup>3</sup> The grouping of items under construct names must be treated with caution, both because this is very preliminary work and because multi-item scales for observational measures of teacher practice are often formative in nature, i.e., items under a construct name do not all “reflect” the construct but can be enacted singly or together to produce the “practice set” which the construct names. See Appendix C.

The third step was to share the draft measures with STEM educators to confirm face validity and obtain feedback on whether the items were comprehensible, comprehensive, and consistent with their practical experience in offering STEM education<sup>0</sup>. Seven STEM educators participated in this round of feedback, which was incorporated into a final draft that resulted in 26 items divided into six new scales: four observational and two based on instructor interviews (see Appendix G).

## Item-level Performance: Inter-rater Reliability

With a draft of the STEM supplement items completed, preparation for piloting the tool in the AfterZone Summer Scholars Program ensued. First, data collectors were

trained in observational note-taking and qualitative coding on the PQA-style rubrics, using videos of STEM programming to simulate program visits. Once raters achieved 80% reliability with pre-determined “gold standard” scores for the videotaped offerings, each was assigned to paired-rater, one-hour observations of at least two instructional sections over the course of the four-week program (see Table 2 for pairing of raters within each offering). Paired raters visited each of the ten program offerings twice, resulting in a total of forty observations, or four observations per program offering. Observers took notes and fit their notes into numeric rubrics to produce scores for program quality. Raters were intentionally “crossed” (e.g., rater pairs were different in most observations) to support estimates of rater bias in future analyses.

**Table 2. Data Collection Design**

	Time 1					Time 2						
	R1	R2	R3	R4	R5	R6	R1	R2	R3	R4	R5	R6
Green Up Summer	X		X				X			X		
Museum of Natural History			X		X			X			X	
Down City Design: Build It		X			X		X	X				
Explore the Bay 1		X		X				X			X	
Explore the Bay 2			X			X			X	X		
Art Explosion	X		X				X				X	
River Adventures				X		X			X	X		
Sailzone				X	X						X	X
Mechatronics		X				X		X		X		
Urban Naturalist	X	X					X			X		

Because this measure is thought to capture formative rather than reflective constructs (see Appendix C for review of formative vs. reflective measure development), maximizing inter-rater reliability at the item level is an important first step. Although raters achieved an acceptable degree of perfect agreement (defined as 80% or higher) with gold standard scores created for video observations during training, this degree of inter-rater reliability was not consistently maintained for all items during live observations. The first column in Table 3 describes the average percentage perfect agreement across all rater pairs during the second wave of 10 paired on-site observations.<sup>5</sup> Raters achieved an average of 80% perfect agreement on seven of the seventeen items. To provide another perspective on variation in scores provided by the rater pairs across both wave of observations, the second column in Table 3 presents intra-class correlations, which, for each item reflect the proportion of score variance occurring within each of the 20 rater pairings to the amount of variance across all raters. Higher ICCs indicate that much less score variance is occurring within the 20 rater pairings (raters agree) than across all 40 individual ratings for that

item – or that paired raters agree. Ratings for eight of the 17 items exhibited acceptable ICCs, defined as those greater than or equal .70 (James, Demaree & Wolf, 1984). Twelve of the seventeen items achieved either 80% perfect agreement or an ICC of .70.<sup>6</sup>

In addition to item-level reliability, all of the STEM items reviewed by expert practitioners were identified as an important part of their instructional practice during most offerings (see the third column in Table 3), indicating a high level of substantive validity of the STEM PQA. The term substantive is used in column 3 to reflect the fact that STEM supplement items are describing practices that instructors are actually trying to produce during the offerings data collectors were observing, indicating that it is possible to see the practice of interest using our method of sampling offerings for observation by external raters. Column 4 describes the % of observed offering sessions in which each of the practices were actually observed by a rater, again suggesting that items on the STEM PQA supplement were actually available to observe during the sampled offering.

<sup>4</sup> Following the study, we vetted the STEM PQA supplement with an expert on STEM instruction research. See Appendix F for comments.

<sup>5</sup> Importantly, these indices of inter-rater agreement were higher than those for the first wave of data collection, indicating that data collectors were becoming more accurate with practice. This is not surprising given that the first wave was the first use of these measures and observer training methods.

<sup>6</sup> When the STEM item scores were dichotomized (1=0, 3=1, 5=1) to indicate simple presence or absence of a practice, the average percent perfect agreement across all 20 rater pairings was 77%.

**Table 3. Item Level Inter-Rater Reliability, Degree of Expectation, and Frequency of Presence in Observed Offerings**

Item	Average % Perfect Agreement Time 2 (N=10 paired raters)	Item Intra-Class Correlations	% STEM Educators Stating Expecting Behavior During Offering Session	% of Observed Offering sessions in Which Practice was Present (N = 20 offerings)
IV.S 1 -More than once, staff help youth connect current activity to personal experiences, application, or previous knowledge.	70	.35	-	95
IV.S 2 -More than once, staff help youth connect current activity to broader societal problems or ethical issues.	50	.72	-	60
IV.S 3 -More than once, staff help youth connect current activity to careers, career preparation, or job-related activities.	50	.68	-	60
IV.S 4-More than once, staff call attention to connections among STEM concepts or disciplines.	50	.76	-	80
STEM.T1 -Staff support youth in identifying a guiding question.	80	.75	-	45
STEM.T2 -Staff support and facilitate youth understanding of steps in the scientific method or STEM design process.	80	.56	100	40
STEM.T3 -Staff ask youth to make predictions, conjectures, or hypotheses.	70	.64	100	80
STEM.T4 -Staff support youth in using a simulation, experiment, or model to answer questions, explore solutions, or test hypotheses.	40	.51	100	65
STEM.T5 -Staff support youth in analyzing data to draw conclusions.	80	.81	100	55
STEM.U1 -Staff support youth in collecting data or measuring.	60	.81	100	90
STEM.U2 -Staff support youth in recording data or observations about events, actions, and objects.	60	.56	100	80
STEM.U3 -Staff support youth in using tools of the field.	90	.81	100	80
STEM.U4 -Staff highlight value of precision and accuracy in measuring, observing, recording, or calculating.	80	.81	100	55
STEM.V1 -Staff model use of STEM vocabulary terms.	80	.46	100	95
STEM.V2 -Staff support and encourage youth in use of STEM vocabulary.	30	.20	86	95
STEM.V3 -Staff support youth in using classification or abstraction, linking concrete examples of principles, laws, categories, or formulas.	80	.73	86	65
STEM.V4 -Staff support youth in conveying STEM concepts through symbols, models, or other nonverbal language.	50	.52	100	70

# Exploration of Multi-Item Constructs and Validity

## Content Validity and Scale Reliability

Content validity of the measure was investigated using dimension reduction techniques in SPSS. Our purpose was to identify items that group together (i.e., a high score on one correlates with a high score on another), thus informing both the content and number of scales maintained in each measure.<sup>7</sup> Specifically, we conducted principal factors extraction with oblimin rotation. Oblimin rotation produces a non-orthogonal oblique solution—that is, it allows extracted factors to correlate, which is appropriate when scales are theoretically correlated with each other. We conducted this analysis using 22 of the 23 original items on the STEM PQA (one interview item was eliminated from the analysis because it had no variation).

Initial analysis suggested the removal of additional items, resulting in the elimination of the items in Scale S (connections) and the two interview question scales, each

of which had low internal consistency<sup>8</sup>. The final principal factors extraction then included the 13 items found within scales T (scientific reasoning), U (observation & measurement), and V (representation) of the STEM supplement (See Appendix G). Within this 13-item analysis, four factors with eigenvalues over Kaiser’s criterion of 1 were extracted, collectively explaining 77.16% of the variance. Table 4 shows the factor loadings after rotation, with loadings below .40 removed from the table. Note that sign is not relevant; rather the size of factor loadings indicates their association with components.

The components may be interpreted as follows. Component one appears to represent the application of STEM Reasoning, component two represents STEM Design, component three represents STEM Data Collection and Measurement, and component four represents use of STEM Vocabulary. Each of these four scales yielded a Cronbach’s alpha coefficient above .70, indicating high internal consistency. Furthermore, the STEM total score – a composite of each of these scales – yielded a Cronbach’s alpha coefficient of .84 and an intra-class correlation of .79.

**Table 4. STEM Factor Analysis Results**

	Component			
	1	2	3	4
T5. Staff support youth in analyzing data to draw conclusions	.90			
V3. Staff support youth in using classification or abstraction	.84			
T3. Staff ask youth to make predictions, conjectures or hypotheses	.82			
T1. Staff support youth in identifying a guiding question	.80			
U2. Staff support youth in recording data or observations	.69		-.56	.59
V4. Staff support youth in conveying STEM concepts through models, symbols or other nonverbal language		.92		
T4. Staff support youth in using a simulation, experiment or model		.90		
U3. Staff support youth in using tools of the field			-.90	
U1. Staff support youth in collecting data or measuring			-.87	
T2. Staff support youth in understanding the scientific method	.66	.45	-.69	
U4. Staff highlight the value of precision and accuracy in measuring, observing, recording or calculating			-.63	
V2. Staff support youth in the use of STEM vocabulary	.46			.92
V1. Staff model use of STEM vocabulary terms				.92
Eigenvalues	5.66	1.88	1.39	1.10
Percent of Variance	44	14	11	8
Reliability: Internal Consistency (Cronbach’s $\alpha$ )	.87	.74	.86	.79
Reliability: Inter-rater (Intra-Class Correlations)	.86	.53	.77	.42
<b>Bi-variate Correlation</b>				
STEM Reasoning		.46	.68	.37
STEM Design			.49	.03
STEM Data Collection & Measurement				.26
STEM Vocabulary				

<sup>7</sup> This is explicitly not an effort to evaluate construct validity in a confirmatory sense, but rather to explore structure in the data and to produce a total score with adequate internal consistency for use in subsequent validation analyses.

<sup>8</sup> Again, these scales are clearly formative in nature and should probably be treated as indexes rather than reflective scales. See Appendix C

The remaining validity analyses in this report utilize these reformulated scale scores. Note that our intent in these analyses is not necessarily to change the originally proposed structure of the STEM supplement but to achieve data reduction (i.e., produce a total score) that could be used in subsequent analyses.

## Concurrent Validity

In order to establish concurrent validity, or the degree to which a new tool compares to a previously established and validated tool that measures similar constructs, we calculated correlations between the STEM supplement total score (average across the four new STEM scales identified in the factor analysis) and the Youth PQA instructional total score. Correlation between the two measures were moderate  $r(20) = .55, p < .05$ .

## Predictive Validity

In an effort to investigate the predictive validity of data from the STEM supplement, we analyzed the extent to which STEM instructional quality (a) predicted absolute levels of child-level measures and (b) predicted change in child-level measures. We employed multilevel models, with youth nested within sites. Specifically, three child outcomes were explored – math efficacy, science efficacy, and possible selves. Math efficacy and science efficacy, or the degree to which youth believe they are “good at” math and science, respectively, were each measured by 12 items on a 4-point Likert scale. Higher scores represent higher levels of efficacy. Possible selves, or the personified cognitive representation of oneself in the future (Markus & Nurius, 1986), and, specifically, career possible selves was assessed by asking youth to what they would like to be when they grow up. A categorical variable was then created by coding their responses as either “0” if the future career possible self was not in the STEM field or a “1” if it was in a STEM field.

The first step was to test whether any significant child-level change in either efficacy or possible selves occurred from the pretest to the posttest. Results indicate that there was no significant difference in the overall sample between math self-efficacy scores at time 1 ( $M = 2.90, SD = .623$ ) compared to time 2 ( $M = 2.91, SD = .550$ ),  $t(110) = -.205, p = .838$ . Likewise, results indicate that there was no significant difference between science self-efficacy scores at time 1 ( $M = 2.85, SD = .612$ ) compared to time 2 ( $M = 2.83, SD = .651$ ),  $t(110) = .459, p = .647$ . Descriptive statistics for change in STEM possible selves yielded similar results, with the majority of youth - 76.8% - reporting no change ( $n = 142$ ), 15.7% ( $n = 29$ ) reporting a decrease in STEM possible selves, and 7.6% ( $n = 14$ ) reporting an increase in STEM possible selves. Even though mean levels

of the child outcomes did not change from baseline to post-test for the entire student sample, it is still possible that mean change scores varied between sites and that this change could be empirically associated with variation in quality as measured by the STEM PQA supplement.

The next step was to investigate between-site variation in the child outcome measures. We used hierarchical linear models to see if enough between-program variance existed to warrant further exploration. To do this, we estimated unconditional models to produce intra-class correlations (ICC), defined as the ratio of site level variance to all variance in the child measures. For both efficacy variables, the ICCs were close to zero, indicating virtually no between-site variance. Therefore no further analyses were conducted with these two dependent variables. However, approximately 13% of the variance in youth-reported possible selves was between sites. As such, hierarchical linear models were used to conduct more in-depth analyses of the Possible Selves outcome data. The model is described in Figure 2 where  $\eta_{ij}$  is defined as the log odds of observing the response (youth endorsing a STEM related Possible Self at Time 2) for youth  $i$  in site  $j$ ;  $\beta_0$  is the intercept in site  $j$ ;  $\beta_1$  is the Time 1 Possible Selves dummy variable;  $\gamma_{01}$  is the site level score on the New Stem Scale of the Youth PQA, and  $\mu_{0j}$  is the error term at Level 2. Because this is binary outcome variable and the Bernoulli method is used to define the model, Level 1 error is not included in the equation

$$\text{Level 1: } \eta_{ij} = \beta_{0j} + \beta_{1j}(\text{PossSelf1})_{ij}$$

$$\text{Level 2: } \beta_0 = \gamma_{00} + \gamma_{01}(\text{NewSTEM})_j + \mu_{0j}$$

**Figure 2. Hierarchical Linear Model (logistic) Used to Estimate an Association Between STEM Instructional Quality and Possible Selves**

In Table 5, one can see that when one more STEM-related staff instructional practices is present, the odds that youth would include a STEM career as a possible self is 2.11 times the odds without that practice present (controlling for possible selves at time 1). In more concrete terms, for the five sites with lowest STEM total scores, an average of 80.5% of youth attending either remained the same in their future self designation or changed from non-STEM to a STEM related career preference. In contrast, for the five sites with the highest STEM total scores 90.0% either remained the same or changed their career preference to a STEM related field. These analyses were conducted with the understanding that sample size was small (185 youth surveys nested in 10 programs; average of 18.5 youth per program), which both limits statistical power to detect

**Table 5. Possible Selves: New STEM**

	Model 1, No pre-test		Model 1, With Pre-test	
	Coefficient (Odds Ratio)	P-values	Coefficient (Odds Ratio)	P-values
Intercept, $\gamma_{00}$	-0.71 (0.49)	0.02	-0.91 (0.40)	<0.01
New STEM Total Score	0.79 (2.21)	0.06	0.75 (2.11)	0.07
Possible Self Baseline (pre-test)	--	--	2.33 (10.31)	0.00

significant relationships and makes the positive finding not generalizable to any larger population. Further, it is almost certain that the model is mis-specified, meaning that critical variables are missing.

## Summary of Findings and Recommendations

All findings provided in this report should be treated with caution, deriving from exploratory analyses. Specifically, conclusions suggested in this study may not be true for other samples and STEM instructional contexts. We offer these caveats based on small sample size, the preliminary state of the measures that comprise the STEM supplement, the relatively untested methods used to guide data collection and observational sampling, and the constraints on the investigation due to limited resources. With these cautions stated at the outset, we summarize the findings from this study as follows:

1. The items on the STEM supplement reflect high quality practice as described in the existing scientific literature on informal STEM programming. These practices were also important in the opinion of the curriculum developers for the AfterZone Summer Scholars program. Curriculum developers reported that these practices would be enacted in most offering sessions and therefore would be available for observers to see (i.e., the measures were reasonably well aligned with the interventions).
2. Many of the items developed to measure STEM instructional practices in the Summer Scholars program demonstrated acceptable levels of inter-rater reliability. Others did not. More work is necessary to improve both the clarity and definition of the practices described by each item, as well as to improve the effectiveness of the data collector training and data collection methodology.
3. While our item development process produced a pool of items that describe high quality instructional practices, all of these items do not perform in reflective manner, i.e., subsets of the items group under a larger category of behavior, or scale, which each item in the scale “reflects.” However, we were able to define several sets of items which operate in this reflective manner. Data reduction techniques (exploratory factor analyses) were used to derive four distinct scales– STEM Knowledge, STEM Methods, Data Collection & Measurement, STEM Vocabulary – each of which yielded Cronbach’s Alpha coefficients above .70, indicating acceptable internal consistency. Further, the STEM total score, constructed as an average across these four categories of STEM instruction, was moderately correlated with the Youth PQA, a generic measure of high quality practice. At this early stage we do not recommend deletion of any items from the STEM supplement because all items describe important practices according to both the

literature and expert review – and these items may serve a useful purpose as a source of individual level performance feedback to practitioners who are observed. However, the four scale structure should be tested again in future replication samples and will likely influence future revisions of the STEM supplement measure.

4. Finally, the STEM total score was associated with child reports that they could envision a “future self” that included membership in a STEM profession. This association was established in a multi-level model which accounted for the nested structure of the data, multiple children sharing the common experience of a Summer Scholars site. Further, this association remained substantively large and statistically significant, even when controlling for student baseline responses, suggesting an association between STEM instructional quality and change in student beliefs.

In summary, we make the following recommendations for use of the STEM supplement:

- The STEM supplement is not appropriate for high stakes uses. Scores on the STEM supplement should not trigger consequences that observed staff experience in a negative way.
- Users of the STEM supplement should read the items carefully and decide for themselves, based on their intended curriculum and reviews of the literature, how well the STEM supplement items “fit” their specific use. If the items on the STEM supplement do not describe the instructional practices they are trying to deliver, it is not the appropriate measure and should not be used.
- The STEM supplement is adequate for purposes of performance feedback to staff in low stakes conditions. Once scores are produced, observed staff should be encouraged to interpret the meaning of STEM supplement scores for themselves. This tool can serve a positive educational purpose and support continuous improvement planning and action.
- The STEM supplement can be used by either external raters or as a program self-assessment. Rudimentary training to improve inter-rater reliability is available from the Weikart Center. As a self-assessment the STEM supplement can be used following guidelines for program self-assessment designed for the Youth PQA.

The Weikart Center will make the STEM supplement to the Youth PQA available to other practitioners and system administrators at no charge and with no assurances regarding reliability and validity of the data which is produced using the tool. Other researchers are encouraged to contact the Weikart Center to gain access to the STEM supplement for use in their studies. In part, this report is designed to offer guidance in those applications.

---

## Works Cited

- Blazevski, J., & Smith, C. (2007). Inter-rater reliability on the youth program quality assessment. Ypsilanti, MI: HighScope Educational Research Foundation.
- Bohnert, A., Fredericks, J., & Randall, E. (2010). Capturing unique dimensions of youth organized activity involvement: Theoretical and methodological considerations. *Review of Educational Research*, 80, 576-610.
- Bollen, K. A. (1984). Multiple indicators: internal consistency or no necessary relationship? *Quality and Quantity*, 18, 377-385.
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology*, Vol. 1. Theoretical models of human development (6th Ed.) (pp. 793-828). New York: Wiley.
- Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, 16, 137-163.
- Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, 17(4), 263-282.
- Durlak, J. A., Weissberg, R. P., & Pachan, M. K. (2010). A Meta-Analysis of After-School Programs That Seek to Promote Personal and Social Skills in Children and Adolescents. *American Journal Community Psychology*, 16.
- Fredericks, J., McCloskey, W., Meli, J., Montrosse, B., Mordica, J., & Mooney, K. (2011). *Measuring student engagement in upper elementary through high school: A description of 21 instruments*. Greensboro, NC: Institute of Education Sciences National Center for Educational Evaluation and Regional Assistance.
- James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group inter-rater reliability with and without response bias. *Journal of Applied Psychology*, 69(1), 85-98.
- Jarvis, C., Mackenzie, S., & Podsakoff, P. (2003). A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. *Journal of Consumer Research*, 30, 199-218.
- Larson, R., & Angus, R. M. (2011). Adolescents' development of skills for agency in youth programs: Learning to think strategically. *Child Development*, 82, 277-294.
- Larson, R., & Brown, J. R. (2007). Emotional development in adolescence: What can be learned from a high school theater program? *Child Development*, 78(4), 1083-1099.
- Larson, R., Hansen, D., & Moneta, G. (2006). Differing profiles of developmental experiences across types of organized youth activities. *Developmental Psychology*, 42, 849-863.
- Lerner, R. M. (2006). Developmental science, developmental systems, and contemporary theories. In R. M. Lerner (Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 1-17). Hoboken, NJ: Wiley.
- Lerner, R. M., Lerner, J. V., Almerigi, J. B., Theokas, C., Phelps, E., Gestsdottir, S., et al. (2005). Positive youth development, participation in community youth development programs, and community contributions for fifth-grade adolescents: Findings for the first wave of the 4-H study of positive youth development. *Journal of Early Adolescence*, 25(1), 17-71.
- Markus, H., & Nurius, P. (1986). Possible Selves. *American Psychologist*, 41(9), 954-969.
- Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, 29, 1-41.
- Schoggen, P. (1989). *Behavior settings: a revision and extension of Roger G. Barker's ecological psychology*. Stanford, CA: Stanford University Press.
- Seidman, E. (in press). An emerging action science of social settings. *American Journal Community Psychology*.
- Smith, C., & Akiva, T. (2008). Quality accountability: Improving fidelity of broad developmentally focused interventions. In H. Yoshikawa & B. Shinn (Eds.), *Transforming Social Settings: Towards Positive Youth Development*: Oxford University Press.
- Smith, C., Akiva, T., Sugar, S. A., Lo, Y.-J., Frank, K. A., Peck, S. C., et al. (2012). Continuous quality improvement in afterschool settings: Impact findings from the Youth Program Quality Intervention study. Ypsilanti, MI: David P. Weikart Center for Youth Program Quality.
- Smith, C., & Hohmann, C. (2005). Full findings from the Youth PQA validation study. Ypsilanti, MI: HighScope Educational Research Foundation.
- Smith, C., Peck, S. J., Denault, A., Blazevski, J., & Akiva, T. (2010). Quality at the point of service: Profiles of practice in afterschool settings. *American Journal of Community Psychology*, 45, 358-369.
- Yohalem, N., Ravindranath, N., Pittman, K., & Evannou, D. (2010). *Insulating the education pipeline to increase postsecondary success*. Washington, DC: Forum for Youth Investment.

## Appendix A – Youth PQA Items

<b>I. Safe environment</b>	<b>III. Interaction</b>
IA1. Emotional climate is positive	IIIL1. Youth get to know each other
IA2. No evidence of bias	IIIL2. Youth exhibit inclusive relationships
IB1. Program space is safe/free of health hazards	IIIL3. Youth identify with the program offering
IB2. Program space is clean/sanitary	IIIL4. Activities publicly acknowledge achievements of youth
IB3. Ventilation/lighting are adequate	IIIM1. Activities carried out in three different groupings
IB4. Temperature is comfortable	IIIM2. Two or more ways to form small groups
IC1. Written emergency procedures in plain view	IIIM3. Each group has a purpose
IC2. Fire extinguisher is accessible/visible	IIIN1. Youth practice group-process skills
IC3. Complete first aid kit is accessible/visible	IIIN2. Youth mentor individuals
IC4. Other appropriate safety/emergency equip	IIIN3. Youth lead a group
IC5. All entrances supervised	IIIO1. Staff share control of the activities with youth
IC6. Access to outdoor space is supervised	IIIO2. Staff provide explanation for expectations, guidelines
ID1. Space allows youth/adults to move freely	<b>IV. Engagement</b>
ID2. Space is suitable for all activities offered	IVP1. Youth make plans for projects/activities
ID3. Furniture is comfortable/sufficient	IVP2. Two or more planning strategies are used
ID4. Physical environment can be modified	IVQ1. Youth make open-ended content choices
IE1. Drinking water is accessible	IVQ2. Youth make open-ended process choices
IE2. Food/drink plentiful and at appropriate times	IVR1. Youth reflect on what they are doing
IE3. Food/drink are healthy	IVR2. Youth reflect in two or more ways
<b>II. Supportive environment</b>	IVR3. Youth make presentations to the whole group
IIF1. Youth are greeted within 15 minutes	IVR4. Staff get feedback on activities
IIF2. Staff use warm tone/respectful language	
IIF3. Staff smile/make eye contact	
IIG1. Session starts/ends within 10 minutes of scheduled time	
IIG2. Materials/supplies are ready	
IIG3. There are enough materials/supplies	
IIG4. Staff explain activities clearly	
IIG5. Appropriate amount of time for activities	
IIH1. Youth engage with materials/ideas with guided practice	
IIH2. Activities will lead to tangible products	
IIH3. Youth talk about what they are doing	
IIH4. Activities balance concrete/abstract	
IIi1. Youth are encouraged to try out new skills	
IIi2. Youth receive support despite imperfect results	
IIJ1. Staff are actively involved with youth	
IIJ2. Staff support contributions of youth	
IIJ3. Staff make frequent use of open-ended questions	
IIK1. Staff approach conflicts in a non-threatening manner	
IIK2. Staff seek input from youth	
IIK3. Staff encourage youth to examine actions/consequences	
IIK4. Staff follow-up with youth involved	

---

## **Appendix B – PASA STEM Community of Practice Guidelines for Inquiry-Based Learning**

Informal science education offers a unique opportunity for after-school providers to use these elements of inquiry-based learning to bridge the relevance and rigor of both in-school and out-of-school learning. These principles are offered to help guide informal science educators in mentoring and being mentored by youth and other educators. They are also meant to guide the development and honest assessment of programs to improve the quality of learning experiences for youth. The hope is that these principles will help educators not only increase the content knowledge of youth, but more importantly, that they will strengthen the critical thinking skills of youth while deepening their interest and engagement in various STEM fields.

### **The Context of Learning:**

1. Youth guide, shape, and lead their own learning in partnership with educators.
2. Educators are facilitators of student learning, not just transmitters of knowledge.
3. Learning is engaging and fun for youth and adults.
4. Youth engage in real world application and problem-solving which are personally meaningful.

### **The Process of Learning:**

5. Learning is deepened by activating prior knowledge about particular topics.
6. Learning is guided by questions from both educators and students.
7. Youth engage in hands-on practice and observation based upon these questions.
8. Youth record information, analyze data, and form conclusions.
9. Youth and educators reflect upon what has been learned.
10. Youth publicly demonstrate and articulate their content and process learning.

---

## 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.<sup>9</sup> 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<sup>10</sup>, 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.

---

<sup>9</sup> For more information, see Bollen (1984); Diamantopoulos and Sigauw (2006); Jarvis, MacKenzie, and Podsakoff (2003)..

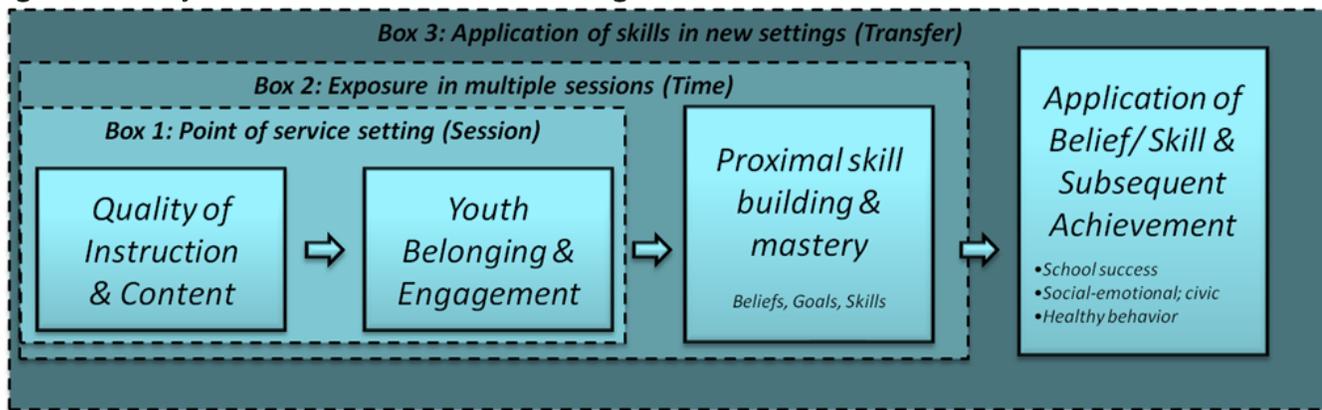
<sup>10</sup> For more information on interpretational confounding, see Hardin, Chang, Fuller and Torkzadeh, 2011.

# Appendix D – Model of OST Contexts and Individual-level Change

Out-of-school time (OST) contexts may play significant roles in the learning and development of young people. OST programs can provide important spaces for positive youth development that, as a young person develops,

complement their growth in school, home, and other contexts (Yohalem, Ravindranath, Pittman, & Evennou, 2010). Figure D1 presents a rudimentary theory of how OST contexts affect individual-level change.

Figure D1: Theory of OST Contexts and Child-level Change



Sustained OST exposure in particular afterschool settings has been associated with positive change in numerous measures of beliefs, goals, and skills: positive self-perceptions, bonding to school, and positive social behavior (Durlak, Weissberg, & Pachan, 2010); emotional regulation (Larson & Brown, 2007); ‘balanced’ possible selves (Oyserman, Terry, & Bybee, 2002); initiative and strategic thinking (Larson & Angus, 2011; Larson, Hansen, & Moneta, 2006). With a few caveats, higher levels of participation (intensity, duration, and breadth) are associated with many of these types of outcomes (Bohnert, Fredericks, & Randall, 2010). However, the specification this research provides about how contexts actually produce individual change, or how individual development emerges through time, in and across settings, can be difficult to generalize into specific practices and curriculum.

Figure 1 presents a likely pathway for youth development and learning in OST settings intended to support practitioners to consider individual change in ways that (1) support intentionality in program planning and delivery and (2) make more efficient use of resources committed to measurement, evaluation and continuous improvement.<sup>11</sup> According to Figure D1, high quality instruction produces youth engagement during a given session (Box 1). The simultaneous presence of high quality instruction and high youth engagement across multiple sessions produces mastery experiences related to the process and content of the sessions (Box 2). Context-specific mastery experiences support longer-term skill development and skill transfer to

external settings, leading ultimately to improved outcomes of interest to policymakers (Box 3). In our model, youth engagement and skill building over multiple sessions mediates the effects of OST setting participation on positive developmental outcomes.

The point-of-service setting is the place where staff, youth, and resources come together as activities (Smith & Akiva, 2008) and is a youth-in-context transactive system (cf. “microsystem” in Bronfenbrenner & Morris [2006]). That is, youth bring their experiences, background, motivation, attitudes, etc., to the point of service, and the setting provides features that include instructional practices and content. Youth engagement over multiple sessions is likely to include regular experience of positive affect, concentration on tasks requiring moderately-difficult effort, and receipt of scaffolding – especially adults’ modeling of the learning task (Fisher & Bidell, 2006) which can be, for example, socioemotional (e.g., using your words), academic (e.g., reducing fractions), or expressive (e.g., design a service project).

<sup>11</sup> The confusing array of names for measures of child and youth outcomes is part of the reason for creating Figure D1, which is designed to help practitioners think clearly about the outcomes they are trying to achieve and empower them to review actual item content, rather than more abstract scale names.

## Appendix E – STEM Literature Review References

A literature review conducted by the Weikart Center identified a number of key elements recommended for out-of-school time programs that promote learning and engagement in science, technology, engineering, and math (STEM). These are largely summarized by the National Research Council's (NRC) (2009, p. 10) recommendations that informal or out-of-school time programs should

1. be designed with specific learning goals in mind (e.g. the strands of science learning)
2. be interactive
3. provide multiple ways for learners to engage with concepts, practices, and phenomena within a particular setting
4. facilitate science learning across multiple settings
5. prompt and support participants to interpret their learning experiences in light of relevant prior knowledge, experiences, and interest
6. support and encourage learners to extend their learning over time

Some of these are assessed by the existing Youth PQA tool as important for all types of youth programs. For instance, interaction, collaboration, active, hands-on learning that utilizes multiple ways for youth to engage, and reflecting on experiences and learning are captured in the Youth PQA, and address key elements # 2 and 3 listed above. The STEM PQA supplement adds an item capturing staff efforts to help youth connect their STEM program activities to prior experiences and knowledge, addressing the NRC's key element #5 above. In addition to connecting to the interests and experiences of youth, the STEM PQA supplement on project-base learning also encourages extended learning over time (#6 above). The STEM PQA Preparation section outlines the expectations that STEM programs will have lesson plans with specific learning objectives, in line with #1 above and asks if the program includes field trips or other ways of connecting youth to multiple STEM-related settings and people (# 4 above). All of these elements foster youth engagement and interest—one of the facets of STEM program which is cited across the literature on STEM practices and learning (Committee on Learning Science in Informal Environments & NRC, 2009; Hussar, et al, 2008; Walker, et al, 2005)

The STEM PQA supplement identifies 3 categories of STEM skills which staff in high quality STEM programs should support: Scientific Reasoning, Observation and Measurement, and Representation. These align with the 3 of the NRC's 6 strands of science learning in informal environments: Understanding Science Knowledge, Engaging in Scientific Reasoning, and Engaging in Scientific Practice. This section of the STEM PQA supplement promotes the key outcomes that various reviews identify as Knowledge, Competence and Reasoning (National Academy of Science ) or Awareness, Knowledge, Understanding (Friedman, 2008). It also captures the juxtaposition of learner's understanding and activities with the formal

abstractions and concepts of the discipline advocated by Fenichel, and Schweingruber (2010).

Another section in the STEM PQA supplement observes whether staff aid youth in making connections between the program activity and a) youth prior experiences and interests, b) societal problems/solutions/ethics, c) careers and career paths, as well as d) making interconnections among STEM concepts and disciplines. Staff support of youth connecting activities to their previous knowledge and interests and potential future careers contributes to youth developing an interest in science (NRC's Strand 1), identifying with the scientific enterprise (NRC strand 6) and career knowledge/acquisition (Hussar, et al, 2008).

The best practices of quality programming identified in the Youth PQA and the STEM PQA supplement are generally appropriate across STEM disciplines. Brunsell (2010) identifies systems thinking, creativity, optimism, collaboration, communication, and attention to ethical considerations as engineering habits of mind which are all supported by staff practices assessed in the Youth PQA and STEM PQA supplement. Youth expressing themselves and reflecting on their experiences—aspects of quality emphasized in the Youth PQA—are facets of quality particularly desired in out-of-school time STEM programming (Brunsell, 2010, Fenichel, & Schweingruber, 2010; Walker, Wahl, and Rivas, 2005). The rubrics and items STEM PQA supplement also describe instructional practices that support key elements of engineering design as delineated by NRC and the National Academy of Engineering (2009) which are: identifying the problem, specifying requirements of the solution, decomposing the system, generating a solution, testing the solution, sketching and visualizing the solution, modeling and analyzing the solution, evaluating alternative solutions, as necessary, and optimizing the final design. As far as out-of-school time programs focusing on mathematics learning, a review of afterschool mathematics literature identified encouraging problem solving, developing and supporting math talk, and emphasizing working together as best practices for math learning in informal settings. (Briggs-Hale, Judd, & Marindill (2006). We believe that staff practices related specifically to mathematics are an area of the STEM PQA supplement that deserve greater attention in the future.

The following sources were consulted in development of the STEM supplement to the Youth PQA:

Ash, D., Klein, C. (2000). Inquiry in the informal learning environment. In Minstrell, J, & vanZee, E.H. (Eds.), *Inquiring into Inquiry Learning and Teaching in Science*. American Association for the Advancement of Science.

[Afterschool Alliance \(2010\). Afterschool and summer programs: Committed partners in STEM education. Retrieved August 18, 2011 from](#)

[http://www.afterschoolalliance.org/STEM\\_JointPositionPaper.pdf](http://www.afterschoolalliance.org/STEM_JointPositionPaper.pdf)

Afterschool Alliance (2010). Afterschool: Middle school science, technology, engineering, and math (STEM). Metlife Foundation Afterschool Alert Issue Brief No. 44. Retrieved from [www.afterschoolalliance.org](http://www.afterschoolalliance.org)

Afterschool Alliance (2011). Afterschool: A Vital Partner in STEM. Retrieved August 2, 2011 from <http://afterschoolpgh.org/news/view/206>.

Afterschool Alliance (2011). Evaluations backgrounder: A summary of formal evaluations of afterschool programs' impact on academics, behavior, safety, and family life. Retrieved August 2, 2011 from [http://www.afterschoolalliance.org/documents/Evaluation\\_sBackgrounder2011.pdf](http://www.afterschoolalliance.org/documents/Evaluation_sBackgrounder2011.pdf)

Beckett, M., Borman, G., Capizzano, J., Parsley, D., Ross, S., Schirm, A., & Taylor, J. (2009). Structuring out-of-school time to improve academic achievement: A practiceguide (NCEE #2009-012). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/wcc/publications/practiceguides>.

Bell, P., Lewenstein, B., Shouse, A.W., & Feder, M.A. Editors (2009). Learning Science in Informal Environments: People, Places, and Pursuits. Committee on Learning Science in Informal Environments, Board on Science Education, National Research Council of the National Academies. National Academies Press. Washington, D.C.

Briggs-Hale, C., Judd, A., Martindill, H., Parsley, D., (2006). Afterschool mathematics practices: A review of supporting literature. Mid-Continent Research for Education and Learning (McREL). Downloaded from [http://www.sedl.org/afterschool/toolkits/math/pdf/math\\_lit\\_rev.pdf](http://www.sedl.org/afterschool/toolkits/math/pdf/math_lit_rev.pdf) or <http://www.mcrel.org>

Brunsell, E. (2011). "[There is an "E" in STEM!](#)" Online posting. 04/25/11. Edutopia.org. [http://www.edutopia.org/blog/science-technology-engineering-math-stem-educationericbrunsell?utm\\_source=feedburner&utm\\_medium=feed&utm\\_campaign=Feed%3A+spiralnotebook+%28Spiral+NoteBook%29](http://www.edutopia.org/blog/science-technology-engineering-math-stem-educationericbrunsell?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+spiralnotebook+%28Spiral+NoteBook%29)

Coalition for Science After School (2007). Science in AfterSchool. New York, NY: The AfterSchool Corporation.

Committee on Learning Science in Informal Environments, National Research Council (2009) Learning Science in Informal Environments: People, Places, and Pursuits. Philip Bell, Bruce Lewenstein, Andrew W. Shouse, and Michael A. Feder, Eds. Washington, DC: The National Academies Press.

Eisenhower Southwest Consortium for the Improvement of Mathematics and Science Teaching (1996). Using community resources to enhance mathematics and science education. Classroom Compass, 3(1).

Falkenberg, K., McClure, P., McComb, E.M. (2006) Science in afterschool literature review. SERVE Center at the University of N. Carolina Greensboro. <http://www.sedl.org/afterschool/toolkits/science/pdf/SERVE%20Science%20in%20Afterschool%20Review.pdf>

Fenichel, M., and Schweingruber, H.A. (2010). Surrounded by Science: Learning Science in Informal Environments. Committee on Learning Science in Informal Environments, Board on Science Education, National Research Council of the National Academies. Washington, DC: The National Academies Press.

Friedman, A. Ed. (2008). Framework for Evaluating Impacts of Informal Science Education Projects [On-line]. (Available at: [http://inisci.org/resources/Eval\\_Framework.pdf](http://inisci.org/resources/Eval_Framework.pdf))

Hall, Georgia, Isreal, Laura. (Sept, 2004). Using Technology to Support Academic Achievement for At-Risk Teens During Out-of-School Time: Literature Review. National Institute on Out-of-School Time for the American Connects Consortium at the Education Development Center, Inc. U. S. Dept of Education

Hussar, K., Schwartz, S., Boisselle, E., & Noam, G.G. (2008). Toward a Systematic Evidence-Base for Science in Out-of-School Time: The Role of Assessment. Cambridge, MA: Program in Education, Afterschool and Resiliency, Harvard University and McLean Hospital.

Institute for Inquiry (n.d.) Pathways to learning: The Institute for Inquiry's approach to teaching and learning science through inquiry Retrieved July 28, 2011 from the website Institute for Inquiry: Examining the art of science education: <http://www.exploratorium.edu/IFI/about/inquiry.html>

Lyon, G. (2011) Lessons learned from 10 years of changing the face of science: New Study validates project explorations youth-science education model. Project Exploration. Retrieved August 2, 2011 from <http://www.projectexploration.org/10years/> Miller, K., Snow, D., Lauer, P. (2004). Noteworthy perspectives: out-of-school time programs for at-risk students. McRel: Aurora California

National Research Council and National Academy of Research (2009). [Engineering in K-12 Education: Understanding the Status and Improving the Prospects](#). Committee on K-12 Engineering Education ; Linda Katehi, Greg Pearson, and Michael Feder, Eds. Washington, DC: The National Academies Press.

National Research Council (2002). Helping Children Learn Mathematics. Mathematics Learning Study Committee, J. Kilpatrick and J. Swafford, Editors. Center for Education, Division of Behavioral and Social Sciences and Education.

---

Washington, DC: National Academy Press. Retrieved from <http://www.nap.edu/catalog/10434.htm>

National Science Board (May 5, 2010). Preparing the Next Generation of STEM Innovators: Identifying and Developing Our Nation's Human Capital, NSB 10-33

Noam G. G. (2010). Youth development, science learning and out-of-school time: The triple alliance. Project Exploration. Retrieved from [www.projectexploration.org](http://www.projectexploration.org)

President's Council on Science and Technology working group (2010). Prepare and inspire: K-12 education in science, technology, engineering, and math (STEM) for American's future [Report to the President] downloaded July 28, 2011 from <http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-stemed-report.pdf>

Project 2061 Benchmarks for Science Literacy. (<http://www.project2061.org/publications/bsl/online/index.php>)

Schwartz, S.E.O., & Noam, G.G. (2007). Informal science learning in afterschool settings: A natural fit? Commissioned paper for the National Academy of Sciences Committee on Learning in Informal Environments. Washington DC.

([http://www7.nationalacademies.org/bose/Schwartz\\_abd\\_Noam\\_Commissioned\\_Paper.pdf](http://www7.nationalacademies.org/bose/Schwartz_abd_Noam_Commissioned_Paper.pdf))

Schwartz, W. (2003) ED478098 2003-06-00 After-school and community technology education programs for low-income families. ERIC Digest

SEDL (n.d.). Goals for Afterschool Learning. Retrieved July 28, 2011 from [http://www.sedl.org/afterschool/toolkits/about\\_toolkits.html](http://www.sedl.org/afterschool/toolkits/about_toolkits.html)

SEDL (n.d.). Principles of Quality Afterschool Science. Retrieved July 28, 2011 from [http://www.sedl.org/afterschool/toolkits/about\\_toolkits.html](http://www.sedl.org/afterschool/toolkits/about_toolkits.html)

Walker, G., Wahl, E., & Rivas, L. M. (2005). NASA and afterschool programs: Connecting to the future. New York: American Museum of Natural History.

Wimer, C., Hull, B., Bouffard, S. M. (2006). Harnessing technology in out-of-school time settings. Out of school Time Evaluation Snapshot, number 7. Harvard Family Research Project [www.gse.harvard.edu/hfrp/projects/afterschool](http://www.gse.harvard.edu/hfrp/projects/afterschool)

---

## Appendix F: Expert Review Comments

This note summarizes the comments and concerns of an expert in the field of afterschool science who reviewed the STEM measure after the pilot had begun. The reviewer highlighted the importance of several YPQA concepts such as teamwork, planning, choosing and reflection in the scientific enterprise and suggested beefing up STEM examples and applications of those concepts. Other areas that could be fleshed out include adding personal observations of the natural world to the connecting to personal experiences item, including in the representation section vocabulary that relates to the scientific approach to learning and knowing such as tentative, firm, corroborative, ambiguous, pattern, and anomalous. The strongest objection was to the “steps of the scientific method”—question, hypothesis, experiment. While the reviewer acknowledged that this is what many people expect, it has long been a source of controversy. “[B]ut this is one of the

science education traps we've been trying to escape for decades. The problem is that the linearity and implied control of the process as defined dramatically limits what science is and therefore the ways that (especially) kids and inexperienced educators can connect to scientific investigations meaningfully. And many hard-nosed natural sciences don't regularly use experiments. Astronomy, most of evolutionary biology, theoretical physics, branches of ecology, and field biology broadly... not to mention the social sciences.” The reviewer recommended including the social sciences and also commented that precision is a double-edged sword—while it is important, intuition, hunches and looking at the big picture in a way that makes sense of the data can be important also, ultimately leading to testable ideas. Another recommendation was including something about appropriate informational resources.

## Appendix G: Original STEM PQA Items and Scales

### S. Staff Connect activities to other learning experiences, issues and applications.

Practice does NOT occur	Practice occurs, but NOT ALL youth/teams engaged	Staff enacts practice AND all youth/teams engaged
1	3	5 More than once, staff help youth connect current activity to personal experiences, applications, or previous knowledge (e.g. "Did you measure angles when building your deck?" "You could use this spreadsheet to record babysitting money").
1	3	5 More than once, staff help youth connect current activity to broader societal problems <u>or</u> ethical issues (e.g. "Can you see that fertilizer run-off contributes to water pollution? Using more solar power, like we did with our go-cart, could reduce dependence on oil. ).
1	3	5 More than once, staff help youth connect current activity to careers, career preparation, or job-related activities (e.g. "There are a number of college majors related to computers.").
1	3	5 More than once, staff call attention to connections among STEM concepts or disciplines (e. g. "An eco-system is affected by geology, climate, and biology: "Understanding angles is also important in physics"; "How might it affect the local wildlife population if engineers removed more pollutants from the water?").
1	3	5 Staff support youth in analyzing data to draw conclusions (e.g., after an experiment, youth are asked to use results to make a generalization, "Your heartbeat increases when you exercise" or "The fastest way to return to the website is to use a bookmark" or "In the survey data we gathered about the election, more people agreed with Candidate A than Candidate B" or "Based on our trials, the large, round tires are best for the rocky terrain.").

---

## T. Staff Support Development of Scientific Reasoning

---

Practice does NOT occur	Practice occurs, but NOT ALL youth/teams engaged	Staff enacts practice AND all youth/teams engaged
1	3	5 Staff support youth in identifying a guiding question (e.g. we want to know if more fish live 300 feet upstream from a sewage pipe than 300 feet below it; how tall is the flag pole in front of the school? How much faster does our computer download if we add x amount of RAM?).
1	3	5 Staff support and facilitate youth understanding of steps of the scientific method (hypothesize, test, conclude, replicate) or STEM design process (analyze, design, implement, test, operate) (e.g. "What steps are there to designing a website?, a bridge made of toothpicks? "If your hypothesis is that smaller birds would eat smaller seeds, first you could test by that putting different size seeds in different feeders, observing, then concluding" "First figure out what you want your robot to do, then design it and test to see if it works correctly" ).
1	3	5 Staff ask youth to make predictions, conjectures or hypotheses (e.g. "if you..., then what will happen?" What will happen if you put baking soda in vinegar? Why do you think it will fizz?).
1	3	5 Staff support youth in using a simulation, experiment or model to answer questions, explore solutions, or test hypotheses (e.g., Youth run a robotics program on a laptop to determine whether it does what they expect it to; Youth try an alternate way to solve an equation and test their result against another example; Youth sample several material types to see which lasts longest when exposed to the elements.).
1	3	5 Staff support youth in analyzing data to draw conclusions (e.g., after an experiment, youth are asked to use results to make a generalization, "Your heartbeat increases when you exercise" or "The fastest way to return to the website is to use a bookmark" or "In the survey data we gathered about the election, more people agreed with Candidate A than Candidate B" or "Based on our trials, the large, round tires are best for the rocky terrain.").

---

---

## U. Staff support in learning observation and measurement skills

---

Practice does NOT occur	Practice occurs, but NOT ALL youth/teams engaged	Staff enacts practice AND all youth/teams engaged
1	3	5 Staff support youth in collecting data or measuring (e.g., Youth use rulers or yardsticks to measure length; Youth use a balance to compare weights; Youth use a thermometer to measure temperature; Youth use a motion sensor to detect movement; Youth count number of different species of birds observed in specified location, number of hits on a website).
1	3	5 Staff support youth in recording data or observations about events, actions and objects (e.g., Youth record temperature changes as a substance is exposed to heat; Youth place markings on the floor to represent the distance traveled by balls of different sizes; Youth place a bent coat hanger on the grass and draw a picture of what they see within it; Youth use a calculator to create a table using an equation).
1	3	5 Staff support youth in using tools of the field (e.g., youth use calculators for mathematics; ph-tests for biology; woodworking tools for building; clay models for design, auto-cad computer design programs).
1		5 Staff highlight value of precision and accuracy in measuring, observing, recording or calculating (e.g. measurement error can impact an experiment or conclusion; a mis-typed letter can bring you to the wrong website, measure twice, cut once, scientists always need to double-check their calculations before drawing conclusions, you must observe carefully to see the difference between species of sparrows).

---

**V. Staff support skills for representing STEM ideas, actions and objects**

Practice does NOT occur	Practice occurs, but NOT ALL youth/teams engaged	Staff enacts practice AND all youth/teams engaged
1	3	5 Staff model use of STEM vocabulary terms (e.g., SCIENCE - chlorophyll, density, atomic, nuclear, geologic, light year, H <sub>2</sub> O; COMPUTERS - hard drive, random access memory (RAM), gigabytes; ENGINEERING -,torque, currents force; MATH - spreadsheet; graph, variable, rate of change, slope, percent).
1	3	5 Staff support and encourage youth in use of STEM vocabulary (e.g. expand upon youth comments with correct terminology; explain meaning of STEM vocabulary in ways youth can understand, ask “do you know the correct term for that?; “ that ‘tall bird’ is a Great Blue Heron”; “saline means it has salt in it.”).
1	3	5 Staff support youth in using classification or abstraction, linking concrete examples to principles, laws, categories, or formulas (e.g. “Mice, porcupines, and squirrels are all rodents, rodents are all mammals.” “The pool ball moved because ‘for every action, there is an equal and opposite reaction’.”).
1	3	5 Staff support youth in conveying STEM concepts through symbols, models, or other nonverbal language (e.g., youth use diagrams, equations, flowcharts, idea webs, outlines, photographs, mock-ups, draft drawings, use of design software to create blueprints, displays, dioramas, physical models, prototypes, graphs, charts, tables, equations, etc.).

---

**Program Preparation: The staff and the program have prepared to maximize STEM learning**

---

1 Staff create lesson plans for almost no STEM activities over the course of the program.	3 Staff create (or will create) lesson plans for some STEM activities over the course of the program.	5 Staff create (or will create) lesson plans for almost all STEM activities over the course of the program.
1 Staff have not identified instructional goals for STEM activities	3 Staff have identified instructional goals for some STEM activities	5 Staff have identified instructional goals for all STEM activities
1 Staff have not linked planned STEM activities to the content of the school day program	3 Staff have linked some planned STEM activities to the content of the school day	5 Staff have linked all planned STEM activities to the content of the school day.
1 Staff have no knowledge of youth academic achievement or challenges	3 Staff have knowledge of some of the youth's academic achievement or challenges	5 Staff have knowledge of all or most of the youth's academic achievement or challenges
1 Safety policies and procedures related to STEM activities have not been established (e.g. internet safety rules, age and supervision guidelines for lab equipment)	3 Safety policies and procedures related to STEM activities have been established, but are not consistently followed (e.g. internet safety rules, age and supervision guidelines for lab equipment)	5 Safety policies and procedures related to STEM activities are established and consistently followed by staff (e.g. internet safety rules, age and supervision guidelines for lab equipment)
1 The program does not expose youth to people or places using STEM (field trips, guest speakers)	3 The program exposes youth to people or places using STEM (field trips, guest speakers) once during a program session, or expose only youth through non-interactive media, such as magazines or videos.	5 Staff expose youth to people or places using STEM (field trips, guest speakers) more than once during a program session.

---

**Project Based: Activities provide continually across sessions**

---

1 None of the STEM activities are part of a multi-session series, or part of a multisession project.	3 At least one STEM activity is part of a series of related activities that span multiple sessions (e.g., one day is evaporation, next day condensation, all tied to the water cycle.)	5 At least one STEM activity is part of a multi-session project (e.g., river cleanup, building a roller coaster using robots, a garden project, etc.)
1 No youth participate in a multi-session STEM project	3 Some youth participate in a multi-session STEM project	5 All youth participate in a multi-session STEM project.

---