High/Scope Youth PQA Technical Report

Full Findings From the Youth PQA Validation Study

Charles Smith and Charles Hohmann
High/Scope Educational Research Foundation

Contents

Summary

The Youth Program Quality Assessment (PQA) is an assessment of best practices in afterschool programs, community organizations, schools, summer programs, and other places where you have fun, work, and learn with adults. The Youth PQA creates understanding and accountability focused on the point of service — where youth and adults come together to coproduce developmental experiences. The ultimate purposes of the Youth PQA are empowering staff to envision optimal programming and building motivation of youth to participate and engage. As an approach to assessment at the systems level, the Youth PQA links accountability to equity by focusing on access to high-quality learning environments for all youth who enroll. As a research tool, the Youth PQA improves measurement of instructional process in places where young people learn.

The Youth PQA consists of seven sections or subscales, each bearing on one dimension of program quality critical for positive youth development: *safe environment, supportive environment, interaction, engagement, youth-centered policies and practices, high expectations,* and *access.* Administration of the Youth PQA employs direct observation of youth program activities for its first four sections and a structured interview with a program director for its remaining three sections. The instrument can be used by outside observers to produce the most precise data or as a program self-assessment directed toward generation of rich conversations among staff.

The Youth PQA Validation Study[1] was a 4-year effort to develop and validate a tool to assess program quality in youth settings. Through the process of instrument development, dozens of expert practitioners and researchers were brought together to provide input on the tool. In total, the validation study encompassed 59 organizations in Michigan and more than 300 Youth PQA observations and interviews conducted in programs serving 1,635 youth. Most of these youth programs were afterschool programs that met weekly or daily over several months. The average age of youth in the sample was 14 years, and more than half were attending programs in an urban context.

The Youth PQA Validation Study employed multiple, independent data sources, including interviews with program administrators, observations in youth work settings, surveys of program youth, expert opinions, and verified reports of staff training. The study's primary concurrent measure of program quality was the Youth Survey from Youth Development Strategies, Inc. All Youth Survey data were independently collected and prepared for analysis by Youth Development Strategies, Inc.

In general, findings from the study demonstrate that the Youth PQA is a valid, reliable, and highly usable measure of youth program quality. Principle findings include:

1. The Youth PQA measurement rubrics are well calibrated for use in a wide range of youth-serving organizations. Average scores fall near the center of the scale and are spread across all five scale points.

[1] The Youth PQA Validation Study was funded by the W. T. Grant Foundation with additional support from the Michigan Department of Education and the Skillman Foundation.

2. Pairs of data collectors were able to achieve acceptable levels of inter-rater reliability on most of the Youth PQA's measurement constructs.

3. The Youth PQA scales subscales are reliable measures of several dimensions of quality. Key subscales demonstrated acceptable levels of internal consistency in two samples.

4. The Youth PQA can be used to assess specific components of programs and is not just a single global quality rating. In repeated factor analyses on two waves of Youth PQA data, the subscales were validated as separate, distinguishable constructs.

5. Youth PQA quality ratings reflect youth reports about their own experiences in the same program offerings. Youth PQA scores demonstrate concurrent validity through significant positive correlation with aligned scores from the YDSI Youth Survey.

6. The Youth PQA measures dimensions of quality that are related to positive outcomes for youth such as youth sense of challenge and growth from the youth program. Youth PQA scores demonstrate predictive validity in multivariate and multilevel models of the data, controlling for youth background variables.

7. Staff in 21st Century afterschool programs find the instrument to have face validity, to be applicable to their current work, and to be a foundation for purposeful change in their programs.

Part I. Introduction to the Youth Program Quality Assessment

The Youth Program Quality Assessment (PQA) is a set of scorable standards for best practice in afterschool programs, community organizations, schools, summer programs, and other places where youth have fun, work, and learn with adults. Using the Youth PQA empowers individuals and organizations to envision optimal quality, to develop a shared language of practice and decision making, and to produce reliable and valid ratings for evaluation and comparison over time. The instrument promotes the creation of environments that tap the most important resource for any youth-serving organization — a young person's motivation to attend and engage. The Youth PQA is based on a multidimensional construct of program *quality at the point* of service where youth, adults, and resources come together to coproduce relational, learning, and adaptive experiences.

The Youth PQA was designed to align with program standards for out-of-school-time organizations and represents an advance in accountability systems design. First, the instrument is designed to support the equity effects of accountability systems because it increases the likelihood that all children who enroll will have access to key developmental experiences — in contrast with outcome-based criteria that provide incentives to attract only high-performing children who will do well on individual measures. Second, the instrument creates the potential for systems to raise the stakes related to quality because it is possible to produce scores with a high degree of accuracy and interpretability. The Youth PQA is an advance over existing quality accountability measures for school-age children and youth that rely upon self-reported checklist or staff self-survey methods, which raise suspicion about bias for numerous reasons. Finally, because the Youth PQA is a dual-purpose instrument, it can reduce the adversarial nature of accountability policies by opening pathways for line staff to use the instrument and for managers to build consensual process around the selection of areas for improvement.

Content of the Quality Construct

At the most general level, the Youth PQA is a quality construct that defines a generic set of experiences that are believed to advance youth development and learning. These characteristics of youth experience can be represented globally as a total score for the instrument's two scales or as a multidimensional construct through individual item and subscale scores. The process of validation tested the instrument's measurement rigor at each step from global scores to subscale scores to items and finally at the indicator level, the instrument's most fine-grained level of measurement. The content and structure of the Youth PQA are detailed in Figure 1. The observation scale consists of subscales I–IV with a content focus on the point of service in youth-serving organizations. Point of service refers to the characteristics of experience that youth have in programs. The observational scales cover issues related to basic safety; staff support; structured interaction and partnerships among program participants; and youth engagement through use of choice, planning, and reflection. These dimensions of quality also describe an implied hierarchy: While many programs can provide adequate safety for children and youth, it is much more challenging to provide higher order experiences such as those described in the interaction and engagement subscales.

The interview scale consists of subscales V–VII with a content focus on practices, policies, and structural resources that organizations use to support quality at the point of service. Although the interview subscales are primarily about characteristics that apply cross-organization, they are focused on the organizational elements that support the production of high-quality youth experiences in programs. The Youth PQA is not a comprehensive measure of quality practices through which

programs extend their reach into nonprogram areas of children's lives such as connection with schools, communities, and families.

These two scales, observation and interview, correspond to the structure of a typical youth-serving organization: offerings within the organization. The quality of youth experiences available in multiple program offerings is captured by completing one or more of the observational ratings. This assessment of offering-level quality is, in turn, nested within the characteristics of a single site or organization that is captured by a single interview rating.

Figure 1. Content of the Youth PQA

Observational Scale – Program Offerings	Interview Scale — Organization
I. Safe Environment	**V. Youth-Centered Policies and Practices**
A. Psychological and emotional safety are promoted.	A. Staff qualifications support a positive youth development focus.
B. The physical environment is safe and healthy for youth.	B. Offerings tap youth content interests to build multiple skills.
C. Appropriate emergency procedures and supplies are present.	C. Youth have influence on setting and activities in the organization.
D. Rooms and furniture accommodate activities.	D. Youth have influence on structure and policy in the organization.
E. Healthy food and drinks are provided.	
II. Supportive Environment	**VI. High Expectations for Youth and Staff**
F. Staff provides a welcoming atmosphere.	E. Organization promotes staff development.
G. Session flow is planned, presented, and paced for youth.	F. Organization promotes supportive social norms.
H. Activities support active engagement.	G. Organization promotes high expectations for youth.
I. Staff support youth to build new skills.	H. Organization is committed to ongoing program improvement.
J. Staff support youth with encouragement.	
K. Staff use youth-centered approaches to reframe conflict.	
III. Interaction	**VII. Access**
L. Youth have opportunities to develop a sense of belonging.	I. Staff availability and longevity support youth-staff relationships.
M. Youth have opportunities to participate in small groups.	J. Schedules are in effect.
N. Youth have opportunities to act as group facilitators and mentors.	K. Barriers to participation are addressed.
O. Youth have opportunities for adult-youth partnership.	L. Organization communicates with families, schools, and organizations.
IV. Engagement	
P. Youth have opportunities to set goals and make plans.	
Q. Youth have opportunities to make choices based on interests.	
R. Youth have opportunities to reflect.	

Assessment Technology

Three core elements of assessment technology differentiate the Youth PQA from checklists and other observation-based quality assessment instruments: a grounded youth development theory, well-honed and tested measurement rubrics, and a systematic observational data collection method. These elements of assessment technology support the dual-purpose nature of the assessment as both a source of precise data about quality and as source of professional development and learning by the staff who use it.

The Youth PQA represents a grounded theory of youth work. The instrument is based on best practices research and a set of youth work methods that have been evolved and tested in 40 years of High/Scope's direct service work with youth. As a result, it has face validity for practitioners because the pieces (items and indicators) seem to fit together in ways that are true to the real circumstances of youth work and education. In our experience of training more than 400 persons to use the instrument, there has been a great deal of comfort and remarkably little resistance to the content of the quality construct. If program staff are going to learn from quality data, the youth program experiences that the data describe must not only be desirable but also possible from the perspective of the user.

A second element of assessment technology is well-honed and well-tested measurement rubrics that allow observers to efficiently capture what happens as young people interact with adults, peers, ideas, and materials. In 2 hours of observation and 1 hour of interview, sufficient data can be collected to score Youth PQA's rubrics with a high degree of objectivity. Each item on the Youth PQA consists of several indicator rows. Each row describes a single program practice in terms of an intuitively simple frequency measure: no children have access to this experience, some children have access to this experience, most children have access to this experience. This simplicity means that the assessment can focus on behaviors that constitute complex adult-youth interactions and still produce reliable and valid data for an entire group of offering participants. Further, because the indicator rows are spelled out in such detail, users can draw substantial support for interpretation of data from the measurement rubrics.

Finally, the Youth PQA is based on a systematic observational method that provides a rare opportunity for program staff to watch what happens in their programs and to watch each other work. The practice of watching and recording what happens in a program, based on a theory of what an optimal youth development ecology looks like, is as important as it is uncommon in the youth work and education professions where staff isolation is often a professional norm. Observation creates contextualized knowledge of program operation and improves interpretation of the generic measures of the youth experience that the instrument describes through quality scores.

Purpose–Methods Continuum

Youth PQA ratings can be generated using different data collection methodologies depending on the ultimate purposes for use of the data. While properties of reliability and validity are inherent in the instrument itself, the neutrality and skill of raters does make a difference to the quality of scores. Scores produced by trained outside observers with known levels of rater reliability produce the most

precise scores. Further, more observation time and multiple ratings are probably a more accurate representation of quality in a complex program than less observation time and fewer ratings. However, when score precision comes at the cost of program staff not buying into the quality improvement process, it may be worth trading the skill and expense of outside observers for less rigorous measurement that produces powerful experiences and insightful conversation for program staff. To meet these competing ends, two data collection methods were developed that provide the ends of a continuum of purpose-driven methods for use of the instrument.

Outside Observer Data Collection Method

In the Youth PQA Validation Study, an outside observer data collection method was employed. For each organization, only one interview was conducted with the site supervisor or program director. However, at each organization three observations were conducted for each of three different program offerings. For example, at one community organization in the study, a complete observational rating was produced for the Double Dutch Club, one complete rating for the Boys to Men club, and one complete rating for the Math Tutoring workshop. Computer classes and hip-hop dance were also part of the programming at this site but were not observed because we limited observation to three offerings per organization. Each 1–2 hours observation of one offering yielded one complete rating on the observational form (subscales I–IV) for that individual offering. The scores for the three offerings were then averaged to create a score that described offering quality at the organization level. Using trained outside observers and a random selection of several program offerings in one organization produces Youth PQA scores with the highest levels of reliability and validity and is the preferred method for use of the instrument when scores of highest psychometric precision are required.

Program Self-Assessment Data Collection Method

The program self-assessment data collection method, by contrast, is a method designed to support powerful dialogue about program quality at lower costs. In the program self-assessment method, program staff complete one composite and consensus-based Youth PQA rating for each site — which may have multiple offerings. Program staff observe each other in 15-minute increments during several different offerings until a sufficient amount of observational evidence is collected to score each item (usually 2–3 hours of total observation time). Staff then meet to assemble the evidence and score a single Youth PQA for the whole site. The interview section is also scored by consensus during an all-staff meeting that includes the director. The self-assessment method may be less reliable than the use of trained external observers for individual offerings, but the end product is a rich conversation among staff about program quality and an opportunity to watch how high-quality experiences are produced.

Figure 2 presents endpoints on a continuum of use for the Youth PQA that ranges between lower stakes use with the program self-assessment data collection method and higher stakes uses that can be achieved through the outside observer data collection method. The total amount of time invested in using Youth PQA will vary substantially with the size of the program and the data collection method used. In general, however, it takes 1–2 hours of observation and 1 hour of scoring for an experienced data collector to complete the four observational subscales for a single, complete observational rating. If multiple observational ratings are produced, then the number of rating should be multiplied by the 2–3 hours required to complete each one. It takes about 1 hour of interview and 1 hour of scoring for an experienced data collector to complete the three interview subscales.

Figure 2. Purpose-Method Continuum for Youth PQA Data

Lower stakes: To get staff thinking & discussing systematically how we are doing and where we want to get better	Purpose (Stakes)	**Higher stakes**: To produce data with maximum precision for monitoring, evaluation, and research
Program self-assessment	Data Collection Method	Outside observer
One PQA per whole site	Number of Ratings	One PQA per offering/classroom
5 hours per site	Time	2 hours per classroom/offering and 2 hours per organization
1 day	Training	3 days
Internal	Audience	External and Internal

Foundations in a Participatory Approach to Learning

The Youth PQA measures a program-quality construct that is rooted in ideas about best practice in youth development, especially an optimal developmental environment focused on learning. The High/Scope participatory approach is a documented and research-validated youth development and learning method (Ilfeld, 1994; Oden, Kelly, & Weikart, 1992; Schweinhart & Smith, 2002; Smith, 2003). The ideas behind the participatory approach are widely shared and described in the language of several disciplines. In general, the approach emphasizes the following:

- Relationships between adults and youth that support the engagement, extension, and transformation of ideas and language and materials.
- Participatory and active learning methods that support effective grouping, leadership, planning, and reflection opportunities that are applied by project for older youth and as a part of the daily routine for younger youth.
- Staff assembly of learning content to support both content and process choices and an appropriate level of challenge for youth.

Emotional well-being, physical health, safety, and nutrition are not explicit parts of the High/Scope work, but they are implicit foundations for the success of the other elements.

The Youth PQA was constructed from this constellation of ideas — safety, relationship, methods, and choice/challenge. Although the items/indicators were vetted with numerous outside experts, the prevalence of these ideas remained central to the quality construct even as the measurement rubrics changed over successive iterations. These core ideas are well represented in the factor structure of the observational portion of the instrument (see Construct Validity: Factor Analyses to Confirm Subscales) and reflect a vision of process quality that fits various models of how people learn: High-quality environments produce positive affect and opportunities for youth to mobilize prior knowledge/experience, which results in intellectual engagement and extension to achieve new understandings and new uses of knowledge (Bransford, Brown, & Cocking, 1999; Marzano, 1998, 2001).

The Youth PQA uses the concept of *learning* to signify effective youth development practice because of an empirical theory that has influenced much of High/Scope's work with children and youth. Put simply, the effects of any intervention are short-lived unless they set in motion an upward cycle of culturally valued behaviors that are linked to incentives for more successful behavior (Berrueta-Clement et al., 1984; Schweinhart et al., 2005). Specifically, if kids learn how to be good learners, they are more likely to be able to adapt to new circumstances when life demands it (to be resilient) and to be rewarded by formal and informal learning systems. In High/Scope's longitudinal studies of both preschool children and adolescents, participatory learning interventions have led to success in subsequent schooling experiences, which in turn led to longer term effects (Oden, Kelly, & Weikart, 1992; Schweinhart et al., 2005).

Part II. Introduction to the Youth PQA Validation Study

The Youth PQA Validation Study is a comprehensive evaluation of the Youth PQA designed to rigorously assess the instrument's reliability, validity, and usefulness. The study was funded by the W. T. Grant Foundation during the period 2001–2005. Although the instrument's structure and measurement rubrics were developed by High/Scope's experienced trainers and research staff, the content of the rubrics was subjected to extensive vetting by expert practitioners and researchers[2] from around the country to ensure face validity for the instrument and interpretability for scores. The instrument review process was facilitated by several organizations including the W. T. Grant Foundation, the Michigan Department of Education, and the Forum for Youth Investment.

Research Questions

The following research questions were of primary interest in the Youth PQA Validation Study:

- Are youth PQA scores reliable? (inter-rater and scale reliability)
- How much observation time is required to produce a stable score for an individual program offering?
- Does the Youth PQA work in a wide variety of program contexts?
- Does the factor structure of the data correspond to the theoretically defined subscales? (construct validity)
- Do Youth PQA scores correspond to what kids say is happening in programs? (concurrent validity)
- Do Youth PQA scores predict positive attitudes and beliefs about program effects? (predictive validity)
- Can youth workers and teachers successfully use the Youth PQA for program improvement? (consequential validity)

Data Sources and Structure

Data Sources

The Youth PQA is designed to measure an ecological phenomenon — patterns of interaction between adults, youth, and resources. Hence, data sources were included that drew information from several elements in the transactional space of a youth-serving program: Youth PQA scores from trained outside observers, Youth PQA scores from interviews with program administrators, surveys of individual youth, expert opinions about program quality, and verified reports on staff training in youth development methods. Several of the findings regarding the performance of the instrument are supported by triangulated evidence from these multiple data sources. Core reliability and validity analyses were replicated in two distinct waves of data collection.

[2] In addition to ongoing review by various stakeholders, data collectors and project staff, three meetings were hosted by the W. T. Grant Foundation in Washington, D.C., on January 27, 2003, December 16–17, 2004, and May 19, 2005. Two additional meetings were hosted by the Michigan Department of Education (MDOE) in October of 2003 and January of 2004 with MDOE and Michigan 21st Century program staff from around the state.

In general, wave 1 data were designed to answer questions about inter-rater reliability and the number of independent observations necessary to produce stable scores. Wave 2 data were designed to answer questions about several dimensions of instrument validity.

Structure of the Youth PQA Validation Study Data

The structure of Youth PQA data is important to an understanding of the research questions, sample design, and data collection used in the project.

The Youth PQA is structured in two levels: youth program offering and youth organization. The offering level refers to specific group-based activities that occur over time with the same staff, same kids, and same purpose. The Youth PQA produces scores at the offering level by scoring observational evidence for each of the quality rubrics that make up the four observational scales (I–IV). In the validation study, the *outside observer data collection method* was employed to collect observational data in program offerings (see description in the Purpose-Methods Continuum section). Where possible, three program offerings were selected for observation, and separate scores (scales I–IV) were generated for each of the three offerings.

The organization level refers to quality dimensions of the whole organization — including all of its offerings — which must be captured through interview because they cannot be efficiently observed. Scores at this level are generated by conducting an interview with a program administrator and scoring this interview evidence on the rubrics that make up three interview subscales (V–VII).

All children 10 years old or above that were currently attending an organization in the study were also surveyed (see the next section). Together these three data sources, Youth PQA organization-level data, Youth PQA offering-level data, and individual youth survey data are assembled into a multilevel data set. Individual children are nested within program offerings, and program offerings are nested within organizations.

The Youth Survey From Youth Development Strategies, Inc.

Survey data from all individual youth-attending organizations in the study were collected using the Youth Survey from Youth Development Strategies, Inc. (YDSI).[3] YDSI is a nationally recognized provider of research, youth assessment, and program quality improvement services. The Youth Survey has been used with thousands of youth subjects in the United States in numerous independent samples and its items and constructs are supported in reports and unpublished analyses available from YDSI. The Youth Survey is a program quality measure that asks students about both the quality of their experiences at a youth-serving organization and about how they feel that these experiences have impacted them. All Youth Survey data were collected and assembled independently of High/Scope by YDSI. All analyses using Youth Survey data were subjected to intensive review by YDSI staff.

For the Validation Study, four subscales from the Youth Survey were selected as most closely aligned with the meaning of the Youth PQA's four observational scales. The Youth Survey subscales

[3] For information about the Youth Survey and YDSI, visit www.ydsi.org.

were *physical safety* (4 items), *belonging* (5 items), *peer knowledge* (2 items), and *decision making* (5 items).[4] A Youth Survey total score was constructed by averaging the four subscales together.

Study Sample

The Validation Study sample includes 59 youth-serving organizations from across the state of Michigan. In total, 356 Youth PQA ratings were completed and surveys were administered to 1,635 youth. The wave 1 sample consisted of data for 335 youth and 22 program offerings nested within 13 organizations. The data for this sample were collected between November 2003 and May 2004. The wave 2 sample consisted of a new group of 1,300 youth and 116 program offerings nested within 46 organizations. The data for the wave 2 sample were collected between May 2004 and March 2005.

The study sample was designed to test the robustness of the instrument for use in many different kinds of program environments where youth are served, primarily, outside of the school day. Because the operational environment for afterschool organizations is so unstable, with organizations emerging, disappearing, or changing auspice, leadership, and purpose frequently, it was impossible to select a stable population from which to draw a random sample. For this reason, the sample was constructed to represent a good mix of the many kinds of youth-serving organizations operating in the metropolitan areas surrounding Detroit and Grand Rapids, Michigan. Selection of offerings within programs was also nonrandom. When organizations were recruited into the study, offerings were selected on the following criteria: each offering must have a regular meeting schedule over at least 3 months with the same group of children for the same general purposes; where possible the three selected program offerings should each have different lead staff.[5]

Characteristics of Organizations

Table 1 describes the types of organizations that participated in the study. The samples for both waves of data collection include a mix of school-based and community-based organizations, located primarily in urban and suburban areas. Most of the organizations accommodated multiple areas of interest and service for their youth rather than a single content focus alone. For those organizations that did focus on a single area of content or service, arts and prevention were the primary foci.

[4] For the wave 2 sample (N = 1,300), Cronbach's alpha coefficients for the four scales were: *physical safety* = 0.56, *belonging* = 0.82, *peer knowledge* = 0.52, decision making = 0.71.

[5] It is likely that this selection method creates selection bias because program administrators were likely to select their best offerings for observation. However, because of the very loose fit between published schedules and actual service delivery, we thought that favored offerings were the most likely to actually occur when scheduled — making data collection more feasible.

Table 1. Description of the Sample for Waves 1 and 2

	13 Organizations Observed in Wave 1	46 Organizations Observed in Wave 2
Program Auspice		
School-Based Organizations (SBO) [includes 21st Century]	Total 23%	Total 35%
Community-Based Organizations (CBO)	Total 77%	Total 57%
Camps	0	8%
Scope of Service		
Single Purpose (SBO + CBO)	15%	20%
Multiple Purpose (SBO + CBO)	85%	80%
Urbanicity		
Rural	0	15%
Suburban	39%	35%
Urban	61%	50%

Characteristics of Program Offerings

Table 2 describes the program offerings that were observed in the two samples. The overall sample captures a wide range of types, purposes, and schedules. The majority of program offerings occurred as part of afterschool programming and met on a weekly basis.

Table 2. Youth PQA Offering Descriptives

	Wave 1 $N = 24$	Wave 2 $N = 118$
Type of Program		
During School	14.3%	6.9%
Afterschool	76.2	53.4
Summer	*	17.2
Residential	4.8	16.4
Other	4.8	6.0
Program Goals *(Do not sum to 100%)*		
Academic Improvement	57.9%	29.9%
Arts/Cultural Enrichment	36.8	55.2
Life Skills Development	57.9	57.5
Summer Employment	5.3	3.4
Service Learning	21.1	27.6
Leadership Development	68.4	*
Other	21.1	14.9
Meeting Frequency		
Less than 1 time per week	15.0%	7.1%
1 time per week	65.0	47.8
2–3 times per week	5.0	4.4
4–5 times per week	15.0	40.7

* = incomplete data

Characteristics of Youth

The study also included a highly diverse youth sample (see Table 3). Over one third of the youth sample was in the 10–12 age range, with the rest of the sample ranging between age 13 and 18. The mean age was 14 years for wave 1 and 13 years for wave 2. The youth sample was evenly divided between Caucasian and African-American youth with 30–40% for each group in both samples. Latino students represented 11.4% and 7.5% respectively for waves 1 and 2. Another 10% were coded as other. Majorities of students in both waves reported moderate to high academic performance with self-reports of B's or better for most of their grades in school.

Table 3. Youth PQA Attendee Descriptives

	Wave 1 *N* = 335	Wave 2 *N* = 1365
Gender		
Male	39.0%	39.5%
Female	61.0	60.5
Age		
10–12	27.6%	37.3%
13–15	38.2	39.6
16 and up	34.2	23.2
Mean age	14	13
Ethnicity		
Black	39.8%	38.0%
Asian	0.8	0.9
Latino	11.4	7.5
Middle Eastern	1.6	2.2
White	32.5	39.9
Other	13.8	11.5
Grades in School		
Mostly A's or A's and B's	54.0%	40.1%
Mostly B's or B's and C's	26.5	26.7
Mostly C's or C's and D's	16.8	9.5
Mostly D's	2.7	3.4

Data Collection

Offering-Level Observations Using the Youth PQA

Between November 2003 and March 2005, Youth PQA data were collected. The wave 1 (November 2003–May 2004) data collection protocol consisted of multiple observations of the same program offering by paired data collectors who observed the program offering at the same time. There were three paired observations for each program offering. The observations typically lasted 1–2 hours for each program offering observed. The wave 2 (June 2004–March 2005) data collection protocol consisted of single observations[6] of program offerings by a sole data collector. The observations typically lasted 1–2 hours for each program offering observed.

Organization-Level Interviews Using the Youth PQA

Youth PQA data for the organization level were collected by first sending each agency administrator a questionnaire that was filled out and returned to the data collectors. The data collectors then conducted hour-long phone interviews with each administrator and recorded answers to neutral questions designed to elicit rich data about the organization's practices and policies, staff development, and strategies for program improvement. The responses to the questionnaire and notes taken during the phone interview were used to score one set of organization-level quality rubrics for each agency that participated in this wave of data.

Youth Data From the YDSI Youth Survey

One person from each agency was trained by YDSI, so the Youth Survey could be administered at the agency. Each site was to survey a minimum of 20 youth ages 10 to 17. Agencies being observed using the multiple-observation protocol for wave 1 administered the Youth Survey in between the first and third observations. Agencies being observed using the 2-hour protocol for wave 2 administered the Youth Survey near the time of the observation(s). Youth Survey data were collected and aggregated independently by YDSI.

Training Data Collectors

From the very beginning of the Youth PQA field studies, we faced the task of providing the data collectors adequate training so that they would understand the instrument and be able to identify the content of specific Youth PQA items in program settings. Our goal was to achieve inter-rater reliability and to develop a training that would make data collectors equally effective, regardless of the data collector's prior education and experience.[7]

The original training model focused on several concrete applications. The first of these is a set of written scenarios that describes events from typical youth programs. These highlight one or another of the dimensions of the Youth PQA. During the training sessions, data collectors read the scenarios and scored the related Youth PQA items. The second aspect of training is the use of videotape clips

[6] The wave 1 data supported the conclusion that only one Youth PQA rating was necessary to produce a high-quality score for any specific program offering so the wave 2 sample was based on the assumption that one Youth PQA rating was a sufficient measure of quality for each program offering.

[7] We checked for data collector bias by examining the relationship between data collector characteristics and Youth PQA scores. The Pearson-*r* correlation coefficients for the Youth PQA total and both data collector age and years of experience working with youth were small and non-significant.

from actual youth program sessions. A third activity asks data collectors to compare item scores from paired raters who gave different scores for the same evidence.

Figure 3. Sample Training Scenario

Sample Scenario
Part I – PROGRAM OFFERING: Youth Opportunities

Item I A: Youth experience psychological and emotional safety in program activities:

Scenario 1 [Score: __/__/]

The group is playing a pick up basketball game. Tony gets pushed and says, "Man, you fouled me." An argument ensues between Tony and Greg. Greg calls Tony a mamma's boy. Tony replies, "At least I am not a faggot like you. I don't dribble and shoot like this. (Tony demonstrates shooting the ball like he thinks a gay person would shoot it.) Tony retaliates by pushing him. Gail, the youth worker, hears the commotion and walks over to see what is going on. Up to this point there has been very little verbal interaction among the players.

Over time all of these materials were refined into a 2-day training model. During day 1, through fitting and scoring of short anecdotes on mailing labels developed from scenarios and film, participants actively engage with every item in the tool in a short amount of time. (See Figure 3 for sample scenario used in a training.) Day 2 focuses more in-depth on taking accurate, objective observation notes and scoring from film or an actual site visit. Throughout the 2nd training day inter-rater comparisons are calculated.

Part III. Primary Findings for Instrument Validation

Findings from the Youth PQA Validation Study presented in this section attempt to answer the research questions presented in Part II. Other findings of significance to the fields of youth development and education will be made available in future reports. These findings examine score distributions and analyses to assess the reliability and validity of Youth PQA scores.

Youth PQA Score Distributions and Sensitivity to Program Change

In order for the Youth PQA to be an effective measurement tool, it must successfully discriminate between program offerings that have higher and lower levels of quality. Practically, this means that the instrument's measurement rubrics must fit the field of practice where quality is being measured. Our very general goal was to develop quality scales that put programs of average quality in the middle of scale for each rubric and where each end of the scale was within the range of potential quality found in our sample of programs. If nearly all youth programs score at the top of the scale, or if the scores cluster tightly in a narrow range, then the instrument will not effectively differentiate among programs. More specifically, it will not capture the variance necessary for researchers to use scores, and is not likely to be of much use to programs that are trying to improve since change will be imperceptible.

Table 4 presents means and score ranges for each of the subscales from both waves of data. The subscale means are simple averages of the scores for items within each subscale (see Figure 1). The min and max scores are subscale scores for the top scoring and bottom scoring programs in each wave of data. These findings suggest that most of the scoring rubrics are well calibrated to the organizations and offerings in this study, with mean scores near the center of the scale (value of three on a five point scale) and scores distributed across most of the scale range. For example, average scores for the Interaction scale in wave 2 was exactly at the middle of the scale (score of 3.03) and scores ranged over nearly four of a possible five scale points (range of 3.83).

Table 4. Score Distributions for Youth PQA Scales

		Wave 1 *N* = 46	Wave 2 *N* = 118
Observational Subscales I–IV			
I. Safe Environment	Mean	4.11	4.40
	SD	.92	.62
	Min	1.00	1.00
	Max	5.00	5.00
II. Supportive Environment	Mean	3.33	3.77
	SD	.85	.83
	Min	1.87	1.68
	Max	4.78	5.00
III. Interaction	Mean	2.74	3.03
	SD	1.03	.90
	Min	1.00	1.00
	Max	5.00	4.83
IV. Engagement	Mean	2.59	2.68
	SD	.99	1.11
	Min	1.00	1.00
	Max	4.67	5.00
Total Score for Subscales I–IV	Mean	3.19	3.47
	SD	.79	.66
	Min	1.63	2.05
	Max	4.49	4.77
		Wave 1	Wave 2 *N* = 36
Interview Subscales V–VII			
V. Youth-Centered Policies & Practice	Mean	*NA*	3.81
	SD		.65
	Min		2.67
	Max		5.00
VI. High Expectations for All Students and Staff	Mean	*NA*	3.77
	SD		.68
	Min		1.67
	Max		4.90
VII. Access	Mean	*NA*	4.43
	SD		.52
	Min		3.00
	Max		5.00
Total Score for Subscales V–VII	Mean		4.01
	SD		.45
	Min		2.62
	Max		4.64

Scores are highest and demonstrate least variance for the safe environment and access subscales (I and VII). These subscales are necessary elements of program quality at both the offering and organizational levels. However, they contain items that are related to both regulatory and socio-cultural baselines for youth serving programs — meaning that most programs are high quality in these areas or supervisors are inclined to report high quality when asked during the interview. In subsequent analyses, poor psychometric performance of subscales I and VII is evident — due to both a lack of variance and poor measurement feasibility for some of the items. The subscales are critical parts of the Youth PQA construct, however, and were retained.

Unfortunately, the Youth PQA Validation Study did not include any consideration of change over time in programs that are undergoing improvement interventions. This is the kind of data necessary to test the sensitivity of the Youth PQA's measurement rubric. However, the Youth PQA was used in one pilot program evaluation of a 21st Century Community Learning Centers (CCLC) afterschool program during the first 3 years of program startup. Youth PQA scores for 16 outside observer ratings at two sites over the 3 years demonstrate steady incremental gains in all 10 items that make up the *safe environment* and *supportive environment* scales.[8] There was no additional quality improvement intervention during this period, suggesting that the instrument is sensitive enough to capture quality improvement as a start-up afterschool program implements over several years. These are also the two subscales that we would expect a start-up program to get better at as staff gain experience and become familiar with regulatory requirements and norms of operation in a school building.

Reliability

Reliability is most easily understood as the consistency with which an instrument produces scores regardless of who is collecting the data or the circumstances of data collection. To assess the reliability of the Youth PQA, analyses were conducted to establish levels of inter-rater reliability and internal consistency on the instrument's seven subscales. Finally, analyses were conducted to establish the amount of observation time necessary to produce a stable score for an individual program offering. The following findings regarding reliability were demonstrated:

- The Youth PQA is an instrument with acceptable levels of inter-rater reliability on subscales II–VI.
- The Youth PQA subscales II–VII demonstrate acceptable levels of internal consistency across two waves of data.
- Youth PQA scores demonstrate a high degree of stability in repeated measures of the same offering.

Evidence of Inter-rater Reliability at the Subscale Level

In the Validation Study, inter-rater reliability was established in two ways: intraclass correlation coefficients were calculated for the seven Youth PQA subscales and the percent agreement for paired raters was calculated for each of the Youth PQA's 30 items. Wave 1 data included three sets of scores from paired raters in each of 22 different offerings.

[8] For the safe environment scale, mean scores were year 1 = 3.61, year 2 = 3.88, year 3 = 4.45. For the supportive environment scale, mean scores were year 1 = 2.43, year 2 = 2.84, year 3 = 3.70.

Intraclass correlation coefficients (ICCs) that were calculated for rater-pairs compare the variance within the rater-pair scores to the variance between all scores for all raters. A high ICC means that there is more variation across all ratings than within individual rater pairs and indicates a high degree of inter-rater agreement. ICCs for this study were estimated using an equation that assumes a random selection of raters across rater pairs (i.e., rater effects are error). ICC greater than .70 are generally considered an indication of acceptable levels of inter-rater agreement, although there is much debate about the use of arbitrary cutoffs for this index (Harvey & Hollander, 2004; James, Demaree, & Wolf, 1984). For example, findings from program quality studies in the medical field suggest that ICCs as low as .40 are interpretable. (See Haut et al., 2002). ICCs for paired-raters on all of the observational subscales were: *safe environment* = .48; *supportive environment* = .69; *interaction* = .83; *engagement* = .72; and *total score* for scales I–IV = .66. All of the ICCs for the observational subscales were in the acceptable range, except for *safe environment.* ICCs for paired-raters (*N* =11 paired ratings) using the three interview subscales were all in the acceptable range: *youth-centered policies and practices* = .51; *high expectations for youth and staff* = .90; *access* = .73; and *total score* for scales V–VII = .80. All of the ICCs for the interview subscales were in the acceptable range, except for *youth-centered policies and practices.*

Evidence of Inter-rater Reliability at the Item Level

For a more easily understood perspective on inter-rater reliability, percent perfect agreement was calculated at the indicator level for 44 rater pairs[9] in wave 1. Percent perfect agreement is the best way to understand rater reliability since more complex statistics such as Kappa statistics and ICCs are influenced by the amount of variance in the sample (Shrout & Fliess, 1979), and we have little other reference about variation in that quality of youth programs to judge the value of a given statistic when it is on the low end. Percent agreement was calculated for each indicator row in the instrument; selection of the same scale point by both raters was considered perfect agreement. Table 5 presents percent perfect agreement at the indicator level for each Youth PQA item using wave 1 data for 44 rater pairs. The first column names the Youth PQA item and the second column lists the number of indicators in the item. The third column presents average percent perfect agreement across all of the indicators for each item. The final column indicates items that had significant revision of indicators between wave 1 and wave 2 data collection.

[9] Although the wave 1 sample had only 22 offerings, each offering was visited three times by a pair of raters and the pairs of raters differed in many cases. We randomly selected two of the rater pairs from each of the 22 offerings for a total of 44 rater pairs.

Table 5. Percent Perfect Agreement at the Indicator Level Using Youth PQA Wave 1 Data

	Number of indicators in the item	Average percent perfect agreement across all indicators in the item	Significant revision in wave 2
I. Safe Environment			
A. Psychological and emotional safety is promoted.	2	80%	
B. Physical environment is safe and free of health hazards.	4	75%	
C. Appropriate emergency procedures and supplies are present.	6	NA	
D. Program space and furniture accommodate activities offered.	4	73%	
E. Healthy food and drinks are provided.	3	NA	
II. Supportive Environment			
F. Staff provide a welcoming atmosphere.	3	71%	
G. Session flow is planned, presented, and paced for youth.	5	58%	X
H. Activities support active engagement.	4	58%	
I. Staff support youth in building new skills.	2	53%	X
J. Staff support youth with encouragement.	3	48%	X
K. Staff use youth-centered approaches to reframe conflict.	4		
III. Interaction			
L. Youth have opportunities to develop a sense of belonging.	4	58%	
M. Youth have opportunities to participate in small groups.	3	72%	
N. Youth have opportunities to act as facilitators & mentors.	3	50%	X
O. Youth have opportunities to partner with adults.	2	50%	X
IV. Engagement			
P. Youth have opportunities to set goals and make plans.	2	65%	X
Q. Youth have opportunities make choices based on interests.	2	60%	X
R. Youth have opportunities to reflect.	4	58%	X

Evidence of Internal Consistency

Internal consistency describes the extent to which items that fall within the same scale are related to each other and capture different aspects of an overarching dimension of quality. To assess internal consistency, Cronbach's alpha was calculated for the seven Youth PQA subscales. Subscales II, III, IV, V, and VI demonstrate acceptable levels of internal consistency and are presented in Table 6. Nunally's (1978) criteria of .70 is widely used as the acceptable standard for scale reliability, although Nunally's earlier work (1967) and research by Davis (1964) view alphas as low as .60 as being acceptable (for a thorough review, see Peterson, 1994).

The safety and access scales are the weak performers for several reasons. First, the measurement rubrics may need improvement. Certain scales top out with scores clustered between scales points 3.5 and 5. Second, the safety scale contains items that do not necessarily vary together even though they may logically be grouped together. This also points out a conceptual problem with scales on most observational measures — the indicators may be the cause of the construct. Most

psychometric operations rely upon the assumption that an enduring state or trait exists that causes observable indicators of the construct to be available for measurement. However, in the observational context, the opposite reasoning may apply. For example, if numerous indicators of the construct safety are in place for several different reasons — different funding requirements, licensing regulations, organizational policies — the indicators may not co-occur in any regular pattern. A program that has more of the indicators might be called safe, but there is little expectation that the various indicators will be highly correlated (see discussion in Bradley, 1999).

Table 6. Internal Consistency for Youth PQA Scales

Observational Subscales I–IV	Wave 1 *N = 22*	Wave 2 *N = 118*
I. Safe Environment (5 items)	.38	.43
II. Supportive Environment (6 items)	.85	.84
III. Interaction (4 items)	.72	.64
IV. Engagement (3 items)	.71	.70
Total Score for Subscales I–IV	.84	.74
Interview Subscales V–VII		*N = 36*
V. Youth-Centered Policies & Practice (4 items)	*NA*	.71
VI. High Expectations for All Students and Staff (4 items)	*NA*	.68
VII. Access (4 items)	*NA*	.45
Total Score for Subscales V–VII	*NA*	.54

Score Stability Between First and Subsequent Observations

Another primary question for the wave 1 data was to establish the number of observations required to produce a stable score for a single offering. Stability refers to the amount of observation time necessary to produce a score that adequately represents the experience available in a specific offering. If the scores change appreciably from one observation period to the next (of the same offering), then multiple observations are necessary to produce an aggregate score that adequately represents the experience that is available to youth in that offering.

Wave 1 data provided an opportunity to examine score stability on both aspects because paired raters visited the same offering three times within a three month timeframe and the offering leader,

group of youth, and purpose of the offering did not change during this period for any of the offerings observed.[10] Offering quality scores were highly consistent in each of three separate observations with Pearson-*r* correlation coefficients ranging from .81 to .98 between the time-one scores and time-two or time-three score.

Construct Validity: Factor Analyses to Confirm Subscales

The meaningfulness of youth data was assessed in several ways in the Validation Study including construct, concurrent, and predictive validity. Construct validity refers to the tendency of Youth PQA item scores to cluster together in patterns that are meaningful because multiple items point to a larger quality concept — or construct — that is being measured. Confirmatory factor analyses are used to determine if the items that were originally clustered as subscales, actually appear in the data when the instrument is used.

Factor analysis is somewhat awkward in application to program quality data, since program environments are fundamentally more malleable and subject to change than the durable mental traits that factor analysis was designed to measure. However, the factor structure of the observational scale generally confirmed our original thoughts about how items should cluster and makes a great deal of intuitive sense as well. Further, across a wide range of program types, average scores for the subscales (factors) take on a pattern that is confirmed in our experience: quality scores decline steadily from safety to staff support and then further to interaction among program participants and finally are lowest on the engagement subscale that measures youth engagement as important decision makers in the program (see Table 4).

The following findings regarding construct validity were demonstrated through confirmatory factor analyses:

- The Youth PQA contains six theoretically derived constructs (subscales) that were confirmed by factor analyses that were replicated on two waves of data.
- The wave 2 version of the instrument yielded a stronger confirmation of theoretically derived factors than the wave 1 version.

Evidence of Construct Validity

Independent factor analyses for two waves of data on the observational subscales are presented in Table 7. These confirmatory models were principal components with varimax rotation. Substantial variance explained by each factor as denoted in the top row of Table 7. In general, the Youth PQA observational subscales are related, but empirically distinguishable constructs, and in each model over 60% or more of the score variance is explained by them. For wave two, a subscale correlation matrix (Pearson-*r*) demonstrates that the three subscales are positively related: II x III = .61; II x IV = .61; III x IV = .62 ($p \leq .01$)

For a cognitive test or other psychological measure, these results might be considered only marginally successful. However, our experience with observational data for program environments

[10] For a full presentation and discussion of analyses conducted to evaluate score stability over multiple observations of the same offering see: Smith, C., Henry, B., & Hohmann, C. (2005, August 20). *Youth Program Quality Assessment Validation Study: Wave 1 findings for reliability and validity analyses.* Report available from the High/Scope Foundation.

suggests that these are successful and interpretable results for this type of data.[11] In both waves of data, three theoretically derived factors were distinguishable. Results for wave 2 are both more promising and more reliable. As described in Tables 1–3, the wave 2 data provided both a more appropriate sample size for factor analysis and a greater variety of program types. The instrument was also improved between the wave 1 and wave 2 data collections, as discussed in Appendix B.

It appears that two items, J and K, in the interaction subscale (see Figure 1), are strongly related to the items in subscale II in both waves of data. We believe that this is due to the fact that these items are influenced by staff behaviors that are the focus of subscale II. However, the intent of these items is more closely focused on behaviors of youth, which should be part of the focus of subscale III. These items have been modified since the wave 2 analysis was conducted, and future research will have to demonstrate improvement of the measurement construct.

Table 8 presents findings for the interview subscales V–VII. Due to substantial revision of the interview scale between the first and second waves of data collection, only findings from wave 2 are presented. Again, the model confirms the presence of related, but empirically distinguishable, factors explaining 64% of total variance for the interview data, with variance explained by the individual factors provided in first row of Table 8. Item I appears to be an independent factor. Subscale correlations (Pearson-*r*) demonstrates that the three subscales are positively related: V x VI = .28, $p = .096$; V x VII = .21, $p = .22$; VI x VII = .36, $p = .03$.

[11] The difficulty for this type of instrument lies in the fact it produces setting scores by mixing together two distinguishable elements — teacher behavior and child response. The benefit is that the Youth PQA captures the moments of interaction in a fairly efficient manner. The cost, we believe, is greater error variance in the item scores. There are almost certainly a cleaner set of "research" constructs in this instrument, i.e., items and indicators could be eliminated to create factors that more clearly differentiate policy-related dimensions of program quality. However, the Validation Study was focused on validation of an instrument that had substantial roots in consensus about best practices, rather than just efficiency and effectiveness as a research tool.

Table 7. Offering-Level Factor Structure for Two Waves of Data Collection

	Wave 1, *N* = 46			Wave 2, *N* = 116		
Percent variance explained by each factor	*26.32*	*25.21*	*16.78*	*29.10*	*18.80*	*12.95*
II. Supportive Environment						
F. Providing a warm and caring atmosphere	0.56		0.52	0.71		
G. Activities are planned, presented, and paced for youth	0.22		0.81	0.67	.52	
H. Providing opportunities to be actively engaged	0.90			0.60		
I. Supporting youth in developing skills	0.78			0.69		
J. Encouraging youth through supportive strategies	0.44	0.59		0.78		
III. Opportunities for Interaction						
L. Helping to develop a sense of belonging	0.58	0.56		0.56		0.12
M. Opportunities to participate in small groups		0.60		0.48		0.84
N. Opportunities for developing leadership skills		0.83				0.65
O. Engaging youth as partners	0.68	0.49			0.54	.34
IV. Engaged Learning						
P. Providing opportunities to set goals and make plans		0.49	0.62		0.71	
Q. Providing opportunities to make choices based on interests		0.43	0.66		0.88	
R. Providing opportunities to reflect		0.73	0.24	0.62	0.37	

Principal components analysis with varimax rotation; 68% of variance explained in wave 1overall; 61% of variance explained in wave 2 overall. All factor loadings below 0.4 deleted, except low loadings on confirmed scales.

Table 8. Organization-Level Factor Structure for Two Waves of Data Collection

	Wave 2, *N* = 36		
Percent variance explained by each factor	*22.39*	*20.24*	*15.62*
	V	**VI**	**VII**
V. Youth-Centered Policies & Practices			
A. Staff qualifications support a positive youth development focus	0.64		
B. Program offerings tap youth interests to build multiple skills	0.54		
C. Opportunities for youth to influence setting and activities	0.87		
D. Opportunities for youth to influence structure and policy	0.73		
VI. High Expectations for All Students and Staff			
E. Staff orientation, meetings, and professional development		0.52	
F. Social norms promote sense of belonging and psychological safety		0.86	
G. Supporting youth in meeting high expectations		0.60	
F. Commitment to ongoing program improvement		0.72	
VII. Access			
I. Staffing patterns promote and sustain youth-staff relationships			
J. Schedules of program sessions are reliable and well publicized			0.79
K. Program locations, schedules, and costs facilitate youth access	0.48		0.55
L. Communicating and collaborating with parents, community, and schools		0.41	0.70

Principal components analysis with varimax rotation; 58% variance explained overall. All factor loadings below 0.4 deleted, except low loadings on confirmed scales.

The safe environment subscale (I) is a part of the Youth PQA's observational scale but is omitted from these factor analyses because inclusion distorts the pattern of item loadings (see note 4). Our plans for future instrument development include treatment of subscale I as a separate observational scale so that the instrument would be constituted from three major scales instead of two (i.e., two observational and one interview).

Concurrent Validity: Youth PQA Scores and Youth Survey Responses

According to the Youth PQA, best practice in youth programming minimally includes safety, supportive adults, youth interaction, and engagement. These attributes are also assessed in the YDSI Youth Survey, a validated measure of youth experiences and attitudes for youth program participants that was administered along with the Youth PQA. Youth PQA scores have the property of concurrent validity to the extent that they are associated with Youth Survey scores for similar kinds of youth experiences, that is, the Youth PQA and Youth Survey are in agreement about which programs are high quality and which are low on dimension of quality that they both measure.

Demonstration of agreement between the Youth PQA scores and Youth Survey scores is a rigorous test of the Youth PQA because the concurrent measure is a completely different data source (a survey of individual youth) and is a direct measure of youth experience in a given offering. The following findings regarding concurrent validity were demonstrated:

- Youth PQA observation subscale scores for *safe environment, supportive environment, interaction,* and *engagement* have moderate to strong correlations with aligned subscales on the Youth Survey across both waves of data.
- The Youth PQA interview subscale score for *youth-centered policies and practices* has moderate to strong correlations with aligned subscales on the Youth Survey across both waves of data.
- The Youth PQA observation *total score* has a moderate to strong correlation with the Youth Survey total score as a global rating of program quality across both waves of data.
- The Youth PQA interview *total score* has a moderate correlation with the Youth survey total score.

Evidence of Concurrent Validity Using the Youth Survey

Table 9 presents Pearson-*r* correlation coefficients as evidence of concurrent validity of the Youth PQA. Individual YDSI subscales were selected as concurrent measures if their content was aligned with a Youth PQA subscale. The Youth Survey total score was constructed as an average of the Youth Survey subscale scores that were aligned with individual Youth PQA subscales and that are listed in Table 9.

For these analyses, sample sizes are substantially smaller than the overall samples of offerings and youth. Youth were included in the concurrent validity sample if they: (1) primarily attend the assessed program offering more frequently than other program offerings available at the organization and (2) attended the assessed program offering frequently or at least sometimes. Program offerings were included in the concurrent validity sample if they included at least five youth who met the youth inclusion criteria listed above.

The bottom of Table 9 presents evidence of concurrent validity for only *Youth-Centered Policies and Practices* subscale. The other two Youth PQA interview subscales, *High Expectations for Youth and Staff* and *Access* do not have close equivalents on the Youth Survey.

Table 9. Concurrent Validity Using YDSI Youth Survey Subscales: Bivariate Coefficients

Youth PQA Observational Scales at the Offering Level	Aligned Scales from the YDSI Youth Survey	Wave 1 *N = 22 Offerings* *N = 129 Youth*	Wave 2 *N* = 29 Offerings *N = 454 Youth*
I. Safe Environment	Physical Safety	.38*	.42*
II. Supportive Environment	Belonging	.44+	.29+
III. Interaction	Peer Knowledge	.69**	.44**
IV. Engagement	Decision Making	.48+	.32*
Youth PQA Total [Scales I–IV]	Youth Survey Total Score	.75**	.47**
Youth PQA Interview Scales at the Organization Level	Aligned Scales from the YDSI Youth Survey	Wave 1	Wave 2 *N = 34 Orgs* *N = 454 Youth*
V. Youth-Centered Policy and Practice	Belonging Interesting Decision Making	NA	.51* .59* .45**
Youth PQA Total [Scales V–VII]	Youth Survey Total Score	NA	.30*

+ = p≤.1; * = p≤.05; ** = p≤.01. Youth Survey total score is a combined score for the following Youth Survey subscales: Physical Safety, Belonging, Peer Knowledge, Decision Making.

Predictive Validity: Multivariate and Multilevel Analyses

One of the most important evaluative criteria for the Youth PQA is how well scores actually predict the youth knowledge/attitudes/behaviors that theory and prior research lead us to believe should follow experiences in high quality programs. Although the YDSI Youth Survey is primarily a measure of youth experiences in a program — and therefore more appropriate for concurrent validity analyses — some of the items do report youth attitudes about the program experience that they attend. For this reason, YDSI Youth Survey scores were modeled as youth outcomes to be predicted by Youth PQA scores in several multivariate and multilevel models.

Several simple hierarchical linear models were estimated to test hypothetical predictive relationships in the data. First, we thought that when a program offering had high scores on the Youth PQA *interaction* subscale — working in small groups, feeling like you belong, facilitating group process, partnering with staff — that the youth in the program would have a sense that they had a greater role in making decisions of consequence in the organization. Second, we thought that when a program had high scores on the Youth PQA *engagement* subscale — goal setting, planning, reflection,

choice — that the kids in the program would have greater sense of interest, challenge, personal growth, and opportunities to give back to their communities. Finally, we estimated a model using the Youth PQA *total score* to predict the YDSI Youth Survey total score. This model is presented here as a general test of Youth PQA data to predict youth outcomes in the form of attitudes about learning.

Findings are presented here only for models that confirm the hypothesized relationships with positive and statistically significant coefficients on the level 2 predictor (the Youth PQA score) in both waves of data. The following findings regarding predictive validity were demonstrated:

- The Youth PQA *interaction* subscale is positively related to youth reports that they *are involved in decision making in the program,* when controlling for gender, minority status, and frequency of attendance.
- The Youth PQA *engagement* subscale is positively related to youth reports that they *are able to giveback to their community* when controlling for gender, minority status, and frequency of attendance.
- The Youth PQA *engagement* subscale is positively related to youth reports that they *experience growth as a result of program attendance* when controlling for gender, minority status and frequency of attendance.
- The Youth PQA *total score* (subscales II–IV) is positively related to the *Youth Survey Total Score* when controlling for gender, minority status, and frequency of attendance.
- The Youth PQA *engagement* subscale was not a consistent and significant predictor of the *youth interest in the program and youth sense of being challenged by the program.*

Evidence of Predictive Validity

Table 10 presents results for several multivariate and multilevel (HLM) models. Each model controls for gender, minority status, and frequency of attendance at level 1 and includes only a Youth PQA subscale or total score at level 2. In each case, substantial variance in the outcome variable exists at level 2 and in each case, the Youth PQA scale explains a substantial amount of total level 2 variance.

Looking at wave 1 analyses, 56 to 82% of the variance among offerings is explained by each respective Youth PQA scale. In addition, with Total Youth Score and Giveback as outcomes, the amount of variance remaining at level 2 after controlling for Youth PQA scale score is insignificant, i.e., the Youth PQA scale predicting each of the model accounts for the variation found between the offerings. While a smaller amount of the variance is explained by each respective Youth PQA scale at wave 2, ranging from 10 to 24%, the amounts still represent a sizable amount of variance accounted for by the Youth PQA scales.

These multivariate, multilevel models are almost certainly underspecified — they lack other important theoretically relevant variables that were not measured in this study. The important point is that Youth PQA scores demonstrate potential for modeling process-outcome relationships where quality scores explain substantial amounts of variance in youth-level data.

Table 10. Predictive Validity Models

		Wave 1		Wave 2
Fixed Effects		**Coefficient**		**Coefficient**
Total Youth Score as Outcome — Youth PQA Total Score				
Intercept		1.01		.12
Youth PQA Total Score		.36**		.14*
Female (male = ref)		-.03		.05
White (minority = ref)		.07		.11
Frequency of Attendance		.05		.05*
Wave 1	**Random Effects:** Intercept	**Variance**: .009	**ICC:** .16	**Variance Explained:** 82
Wave 2	**Random Effects:** Intercept	**Variance**: .029**	**ICC:** .15	**Variance Explained:** 19
Giveback as Outcome — Youth PQA Engaged Learning				
Intercept		1.28		.34
Youth PQA Total Score		.23*		.14+
Female (male = ref)		-.06		.01
White (minority = ref)		.27		.07
Frequency of Attendance		.08		.06
Wave 1	**Random Effects:** Intercept	**Variance**: .019	**ICC:** .09	**Variance Explained:** 56
Wave 2	**Random Effects:** Intercept	**Variance**: .09**	**ICC:** .17	**Variance Explained:** 10
Youth Growth as Outcome — Youth PQA Engaged Learning				
Intercept		1.19		.17
Youth PQA Total Score		.32**		.09+
Female (male = ref)		-.19		-.03
White (minority = ref)		.17		.03
Frequency of Attendance		.06		.06+
Wave 1	**Random Effects:** Intercept	**Variance**: .046*	**ICC:** .20	**Variance Explained:** 63
Wave 2	**Random Effects:** Intercept	**Variance**: .025*	**ICC:** .06	**Variance Explained:** 24
Youth Decision Making as Outcome — Youth PQA Interaction Opportunities				
Intercept		1.02		-.15
Youth PQA Total Score		.29**		.09*
Female (male = ref)		-.26*		.15*
White (minority = ref)		-.18*		.11
Frequency of Attendance		-.07+		.05
Wave 1	**Random Effects:** Intercept	**Variance**: .030*	**ICC:** .21	**Variance Explained:** 77
Wave 2	**Random Effects:** Intercept	**Variance**: .022*	**ICC:** .07	**Variance Explained:** 22

*+ = $p \leq .1$; * = $p \leq .05$; ** = $p \leq .01$*

Global Quality Scores: Expert Opinion and Youth Worker Training

So far validity analyses have focused on Youth PQA subscales as powerful independent dimensions of quality that are reliable and valid for use in a wide range of youth development and learning contexts. However, the Youth PQA can be treated as a global quality score by simply averaging the subscale scores together. Tables 4, 6, 9, 10, and 12 all include descriptions of the Youth PQA *total score*.

Additional concurrent validity analyses included establishing the strength of association between the Youth PQA total score and two other global measures of program quality: (1) expert opinion about the quality of organizations in the study and (2) participation in youth development training that is focused on the youth development by offering leaders in the study. The following additional findings regarding concurrent validity were demonstrated:

- The Youth PQA observation *total score* was positively associated with expert ratings of youth-centeredness and availability of resources.
- The Youth PQA observation *total score* was positively associated with offering leaders who had received training in a youth development method.

Evidence of Association Between Quality Scores and Expert Global Quality Ratings

The Youth PQA was designed as a measure of best practice. Two of the key global indicators of best practice in youth programming are use of a youth-centered program philosophy and availability of sufficient resources. If the Youth PQA is valid as an assessment of best practice, Youth PQA scores should differentiate between programs that are high vs. low on both of these dimensions of the best practice.

Spearman's rho coefficients were calculated for the total score on the Youth PQA observational scale (subscales I–IV) and expert ratings on two global dimensions of quality for the 13 organizations in the wave 1 data collection: the organization's youth centeredness and the organization's level of resources. Expert ratings were generated by averaging multiple ratings from seven expert raters over the 13 organizations. Youth centeredness was assessed on the following metric: 1 = Youth centered, 0 = Adult Control or Laissez-faire. Availability of resources used the following metric: 3 = High, 2 = Moderate, 1 = Low. Correlations between the Youth PQA observation *total score* and the expert ratings were 0.59 for youth-centered program philosophy and 0.52 for availability of resources and both were statistically significant at the $p = 0.01$ level.

Evidence of Association Between Quality Scores and Youth Worker Training

Staff training in an explicit youth development approach is an important element contributing to best practice in youth programming. If the Youth PQA is a valid assessment of best practice, it should differentiate between the performances of youth workers who have had training on an explicit youth development approach from those who have not.

Analysis of Variance was employed to test for differences on the Youth PQA observation *total score* (subscales I–IV) for trained and untrained offering leaders. An offering leader was considered

trained if he or she had attended 4 or more days of training on the High/Scope youth-level approach. An offering leader was considered not trained if he or she had not attended any High/Scope training. In the wave 1 sample of 8 trained offering leaders and 14 untrained offering leaders, the mean difference in the Youth PQA *total score* was 1.4 scale points higher for the trained youth workers ($F = 5.25$, $p<.01$). For the wave 2 sample of 37 trained offering leaders and 80 untrained offering leaders, the Youth PQA total score mean difference was 0.4 scale points higher for trained youth workers ($F = 9.66$, $P<.01$).

Consequential Validity

Consequential validation of an assessment tool requires evidence that use of the tool actually supports the attainment of the broader outcomes that the program was designed to accomplish. An example of consequential validation would be evidence that children in a reading program have increased literacy ability at the end of the program because of the use of a diagnostic literacy assessment tool. With the Youth PQA, we were specifically interested in validating an assessment instrument that would bring powerful measurement rubrics and data collection methods into everyday use by youth workers and teachers. The consequences of use — users are able to complete it, put the data to use, and generally like the process — are of great importance to our thinking about validation.

For the Youth PQA, we do not have direct evidence that use of the tool results in better outcomes for youth. We do, however, have evidence that suggests that use of the Youth PQA effectively supports the process of data-driven program improvement. Although not the ultimate purpose of most youth programming, the ability to change youth work practice for the better is certainly one of the key intermediate outcomes (consequences) that the Youth PQA was designed to achieve.

Use of the Youth PQA as Part of an Afterschool Program Evaluation

As a test for consequential validity, we used the Youth PQA as part of a program evaluation for a multisite 21st Century afterschool program during the first 3 years of program operation. Data were collected by trained outside observers and presented to staff following each program year. During the meetings, we facilitated a very generic process of figuring out what to do as a result of the Youth PQA data — with no other explicit program improvement process or intervention in mind. From this experience, several results followed as a consequence of using the Youth PQA:

- The Youth PQA provided a useful framework for conversations about (1) the distribution of program resources and (2) the quality of program processes.
- Using Youth PQA data staff developed logic models that pointed to specific plans and goals for program change.

In staff evaluations of the program improvement sessions over 3 years, five themes were prominent:

- The process helped clarify thinking.
- The process created better communication across groups.
- Participants wanted more time.
- Participants thought that this should be a regular part of their work.

- Participants thought that Youth PQA data should be presented at the beginning of the program year rather than just at the end.

Responses From Afterschool Staff Using the Program Self-Assessment Method

Another source of information related to how users view the Youth PQA comes from follow-up interviews with program directors who assembled staff teams to use the instrument at 21st Century sites during the 2004–2005 program year. Interview comments are included in Table 11.

Customer Satisfaction Ratings for the Youth PQA 1-Day Training

A final source of information regarding how youth workers and teachers responded to the instrument are satisfaction ratings from end-of-training surveys given to Youth PQA training participants. One hundred and fifty-three trainees provided satisfaction ratings on a five-point scale where 1 = poor and 5 = excellent. Trainees overwhelmingly said that the Youth PQA was "applicable to current work" (mean score = 4.45) and that they were engaged by the instrument through the "organization of the training" (mean score = 4.33) and the "level of participation during the training" (mean score = 4.33).

Table 11. Comments From Interviews With 21st CCLC Staff in the Michigan Pilot

Excerpts from 6 interviews	
What did you like about it?	The enlightenment that the staff experienced. "Oh my gosh! I never thought about that!" Their eyes were opened to some areas they had not thought of. (Staff reported that many of the program staff are paid minimum wage and have little formal education related to youth work.) Good conversations came from the process. "The Youth PQA captures what she (local evaluator) has learned in her gut over 35 years. The Youth PQA gives you the common language and position to have the discussions for improvement. It forces the difficult discussions that administrators usually ignore instead of address (e.g., Adult sits in chair and watches youth-no interaction occurs)." The administrators of the school were able to observe the program and learn things they could not have known from hearsay or assumptions. It "helped them with giving them a perspective on program happenings." As a way of presenting the Youth PQA as a positive, low-key way to evaluate their programs and work towards improvement program, director and four staff members embarked upon a "mock Youth PQA." In January, the group met and looked through the Youth PQA and then went back to their sites and collected a couple dozen anecdotes during the month. At the end of the month they met for a "retreat." The collected anecdotes were posted into items and then the group collectively scored the mock Youth PQA. Discussions included why the anecdotes fit and how the scoring was processed. This prompted many in-depth conversations and many insights were encouraging. The approach of the Youth PQA gave them a way to talk to program staff about thinking and planning programs — all details of the planning process.
What changed as a result (e.g., people's perceptions about things and/or specific actions that you undertook as a result of the data)?	The enlightenment of the staff to the many different areas of a self-evaluation/assessment. The elementary site coordinator's perception changed right away. She caught on immediately and made positive and major changes—a major impact on the program. Local evaluator believes that the site coordinator will continue to make changes. - Evaluator observed the elementary program and found "chaos" — youth were sitting and waiting for adults. They had nothing to do and they were loud and physical. That high level of energy was taken into the program activities. - She targeted (1) Engagement—there was nothing for them to do while they waited for the adults to start the activities; and (2) Choices — what options did they have other than choosing to annoy each other? - The Elementary Site Coordinator purchased materials and created a system for snacks and activities. The choices completely turned around the situation. The youth are able to make choices immediately and independently — without waiting for an adult to offer, or start, the process. Because they were actively engaged in activities of their choice they did not have time to get loud and physical with each other. The site coordinator and principals used the Youth PQA findings to discuss and review whether the provider used was good for the youth at their schools. They believe that the provider is not up to par. They plan to/want to use the Youth PQA to build staff development for the next year. The local evaluator put together a presentation and there were many constructive conversations between the local evaluator and staff with regard to strengths and weaknesses — all geared towards improvement. Director believes that the Youth PQA "is a useful tool for really understanding their program quality and how to have conversations to improve it."

Part IV. Capturing Quality at the Point of Service

Representing Point-of-Service Quality for an Entire Organization

Although the Youth PQA has performed with acceptable psychometric precision in the several tests described in this document, a persistent issue remains. What is the best way to represent point of service quality — offering level quality in the language of this report — at the organization level? This is an important question for both accountability and research uses of the instrument.

Youth experience in programs is composite, made up of experiences with different groups of youth, different staff, and for different purposes. Data from the Validation Study show that quality scores vary as much or more among different offerings (different staff) in the same organization as they do across different types of organization. If this is the case, it is likely that a single observational rating that is valid for a single offering will not actually represent overall quality in an organization. Indeed it is a fair question, whether or not organization-level quality as an average score really exists at all — if quality is understood as the quality of youth experiences that are available in an organization — since quality will be determined by the different experiential pathways that individual youth and their families choose as they select offerings for participation from multiple choices available. Regardless of how this question is answered, however, accountability requires program quality scores for organizations. This means that we need to use our best judgment about how to represent point-of-service quality using the Youth PQA.

Table 12 presents Pearson-*r* correlation coefficients for the Youth PQA offering total score and YDSI Youth Survey total scores. Moving from left to right, the cells of the table present correlation coefficients for samples of youth that are less closely nested within to program offerings that were assessed on the Youth PQA during the study. In column one, the relationship between Youth PQA scores and youth reports of experiences at the program are strong. These are single-purpose organizations so an answer about the organization on the Youth Survey refers only to the offering that was assessed by the Youth PQA.

In column two, the strength of the relationship declines but remains in the moderate to strong range. Here the YDSI Youth Survey responses came only from children who primarily attended only the assessed offering frequently, even though they may have spent some time at the organization doing other things.

The third column presents coefficients for Youth PQA and YDSI Youth Survey scores aggregated to the organization level, that is, several offering-level Youth PQA ratings as a point of service quality average score for the entire organization. Here, YDSI Youth Survey responses from any youth *attending any of the assessed offerings* with any degree of frequency are included in the quality score generated from the YDSI Youth Survey.

Finally, in column four, several Youth PQA ratings are again averaged to create an organization level score for point-of-service quality, but only youth who did not attend any of the assessed offerings are included.

Table 12. Correlation Coefficients for Youth PQA Total Score and YDSI Youth Survey Total Score at Different Levels of Aggregation

	Single purpose orgs, all kids attend a single offering	Offering by participant in only that offering	Org. level — All offerings, All kids that attended any of the rated offerings	Org. level — All offerings, Kids who did not attend any of the rated offerings
Wave one	NA	.75* *N =13 offerings* *N =129 kids*	.48+ *N = 9 organizations* *N = 234 youth*	*Awaiting final analyses*
Wave two	.76* *N =8 offerings* *N = 65 youth*	.47** *N = 29 offerings* *N = 454 youth*	.25* *N = 36 organizations* *N = 982 youth*	-.17 *N = 36 organizations* *N = 365 kids*
*+ = p≤.1; * = p≤.05; ** = p≤.01*				

In Table 12, it is the level of aggregation used in column three that is the most obvious method of representing point of service quality for an entire site or organization. The sample size of the wave two coefficient in column three is 118 offerings and 982 children aggregated up to 36 organizations. However, the coefficient is of a much lower magnitude than in columns one and two where youth experiences are more tightly coupled to point-of-service experiences.

What Table 12 suggests is that quality varies dramatically within organizations and that it is critical to be conscious about how well point-of-service ratings represent the totality of experiences available in an organization. In general, several strategies for being more consciously representative are available when using the outside observer method:

- More individual ratings of program offerings are probably better.
- Random sampling of the program offerings where observational rating will be completed may be useful.
- Staff who spend the most time working with youth or who work with the largest number of youth in the organization must get rated.

Instrument Performance Using the Program Self-Assessment Method in 21st Century Afterschool Programs

The program self-assessment data collection method is one way to deal with quality variance at the point of service. In this method, the explicit point is to capture an organization-wide rating for quality at the point of service. As a part of a pilot study for the Michigan Department of Education 21st Century Community Learning Centers, the Youth PQA was completed at 22 sites around the state by teams of staff (for details, see Smith, 2005). This method is designed to capture a totality of program experience at a lower level of psychometric rigor than the outside observer method used in the Validation Study.

Instrument performance in this context was expected to differ from data collected using the outside observer method in two ways: scores would be higher and have more error variance as demonstrated by lower alpha coefficients for the scales. However, instrument performance was better than anticipated. For our sample of twenty-two 21st Century programs — data collectors with only 1 day of training — subscale scores were about one-quarter to three-quarters of a scale point higher than the Validation Study sample. Alpha coefficients for the pilot sample were nearly identical to the Validation Study for the observation subscales (.76 for the observation total score) and lower for the interview subscales (.50 for the interview total score).

Part V. Conclusion

This report reflects results from a comprehensive multi-year development project and validation study for the Youth Program Quality Assessment. Although psychometric evaluation of the Youth PQA suggests that the instrument has some areas of weakness, overall performance of the instrument demonstrates an adequate level of precision for use in both research and applied circumstances. Due to the iterative nature of development and evaluation, we believe that some areas of weakness in the instrument have already been improved but are not reflected in this report since we have not had the opportunity to fully test subsequent changes in the tool. Appendix A describes the iterative process of instrument development through the measurement constructs that were used in wave 2 of this report.

The development of the Youth PQA's measurement constructs, subscales, items, and indicators was driven by both theory and extensive practitioner experience. However, the next stage of evaluation for the instrument will be more fully empirical since new waves of data collection continue to increase the size and variety of samples that we have available. Specifically, larger samples will allow us to more fully test and evaluate the item-level constructs — that is, how well the groupings of indicators that make up each item actually hang together as empirically derived clusters of staff behavior.

Our goal for the Youth PQA Validation Study was to create an inexpensive instrument that focused on the elements of program practice that are most likely to influence youth behavior and outcomes — the point of service where adults and youth come together to co-create program experiences. We hope that the Youth PQA will help to initiate a discussion among researchers, practitioners, and decision makers about how programs should prioritize regulation and resources in the out-of-school time field. This instrument does not provide a single answer to these questions of how to allocate time and effort but we hope that it does present the opportunity for many different kinds of programs to create their own answers to these questions. We hope that others in the field will find the instrument useful, and we encourage others who may be interested in using the tool to seek our support for these efforts.

References

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). *Changed lives: The effects of the Perry Preschool Program on youths through age 19.* Monographs of the High/Scope Educational Research Foundation, 8. Ypsilanti, MI: High/Scope Press.

Bradley, R. H. (1999). The home environment. In S. L. Friedman & T. D. Wachs (Eds.), *Measuring environment across the life span: Emerging methods and concepts* (pp. 31–58). Washington, D.C.: American Psychological Association.

Bransford, J. D., Brown, A. L., & Cocking, R. (Eds.) (1999). *How people learn: Brain, mind experience and school.* Committee on Behavioral and Social Sciences Education, National Research Council. Washington, D.C.: National Academy Press.

Davis, F. B. (1964). *Educational measurements and their interpretation.* Belmont, CA: Wadsworth.

Epstein, A. S. (1993). *Training for quality: Improving early childhood programs through systematic inservice training.* Ypsilanti, MI: High/Scope Press.

Epstein, A. S. (1999). Pathways to quality in Head Start, public schools and private non-profit early childhood programs. *Journal of Research in Childhood Education, 13*(2), 101–119.

Harvey, R., & Hollander, E. (2004, April). *Benchmarking r_{wg} interrater agreement indices: Let's drop the .70 rule-of-thumb.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Haut, S. R., Berg, A. T., Shinnar, S., Cohen, H. W., Bazil, C. W., & Sperling, M. R. (2002). Interrater reliability among epilepsy centers: Multi-center study of epilepsy surgery. *Epilepsia, 43*(11), 1396–1401.

High/Scope Educational Research Foundation. (2003). *High/Scope program quality assessment – Preschool version* (2nd ed.). Ypsilanti, MI: High/Scope Press.

Ilfeld, E. (1994). *Learning comes to life.* Ypsilanti, MI: High/Scope Press.

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.

Marzano, R. J. (1998). *A theory based meta-analysis of research on instruction.* Aurora, CO: Mid-continent Regional Educational Laboratory.

Marzano, R. J. (2001). *Designing a new taxonomy of educational objectives.* Thousand Oaks, CA: Corwin Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Oden, S., Kelly, M. A., & Weikart, D. (1992). *Challenging the potential: Programs for talented disadvantaged youth.* Ypsilanti, MI: High/Scope Press.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *The Journal of Consumer Research, 21*(2), 381–391.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40.* Monographs of the High/Scope Educational Research Foundation, 14. Ypsilanti, MI: High/Scope Press.

Schweinhart, L. J., & Smith, C. (2002). *Effects of recent High/Scope curriculum support on school achievement and discipline referrals.* Report to the North Central Education Laboratory, Washington, D.C.

Shrout, P. E., & Fleis, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–427.

Smith, C. (2005). *Measuring quality in Michigan's 21st Century afterschool programs: The Youth PQA self-assessment pilot study.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Smith, C. (2003). *A pedagogy for civic education: Experiential learning, diversity, and shared control in the Youth Urban Agenda.* Paper presented at the International Civic Education Research Conference, New Orleans, LA.

Appendix: Narrative Description of Instrument Development 2002–2005

The High/Scope Foundation has developed a variety of observational assessment tools over the past four decades. Through the 1970s and 1980s, High/Scope Program Implementation Profiles (PIPs) were developed to track the fidelity of implementation of the High/Scope Approach in preschool, elementary, and secondary school classrooms following staff training. The Foundation's Program Quality Assessment (PQA) instruments are an extension of these earlier efforts toward more generic ideas about best practice in the field, not just in programs using the High/Scope Approach. In addition to the Youth PQA, High/Scope has developed PQAs for use in settings that serve infants and toddlers, preschool children, students in elementary school day and afterschool settings, and children and youth in K–12 special education programs. The Foundation's Preschool PQA was validated in several studies (for use in Head Start programs see Epstein, 1993, 1999; for review of studies in school-based preschool programs see High/Scope Educational Research Foundation, 2003).

Through the 1990s a youth-level PQA was in development for use in community programs and alternative high schools. During the period 2002–2005, the Youth PQA Validation Study was funded by the W. T. Grant Foundation and produced the instrument and validation research reported here. Other important collaborators in the development of the Youth PQA include the Skillman Foundation (2002–2004), the Michigan Department of Education (2003–2005), the Detroit Youth Sports & Recreation Division, Prime Time of Palm Beach County Florida, and the school district and City of Grand Rapids Michigan (2004–2005).

By the spring of 2002, we had completed a first draft of the youth process quality instrument in preparation for a consensus building meeting sponsored by the W. T. Grant Foundation in Baltimore that spring. The consensus building meeting brought together representatives of youth serving agencies throughout the country, many of whom had agreed to serve on the advisory panel of the Youth PQA development project.

Figure 4. Youth PQA July 2002 Item Tally

PROGRAM SETTING: *Opportunities and Supports*
Safe and healthy physical environment for youth.
Emergency procedures and supplies to ensure the safety of youth.
Setting accommodates a variety of activities.
Materials are organized and accessible.
Materials and technology are appropriate to activities.

PROGRAM PROCESSES: *Youth Opportunities*
Opportunities to develop and experience a sense of community.
Opportunities to engage in active learning.
Opportunities to work cooperatively in groups.
Opportunities to plan and share their plans.
Opportunities to make choices based on interests.
Opportunities to reflect.
Opportunities for leadership.
Opportunities to practice communication skills

PROGRAM PROCESSES: *Youth Supports*
Warm and caring atmosphere.
Activities planned, presented, and paced for youth.
Support in learning or using new skills.
Variety of encouragement strategies.
Staff and youth as partners.
Staff as positive role models and mentors.
Youth-centered approach to resolving conflicts.
Staff support youth in defining and achieving success.

Feedback on the Youth PQA from the participants in the consensus building meeting provided the basis for further revisions in the instrument that were undertaken through the spring and early summer of 2002. The resulting version was identified by its date of completion — the July 2002 version — which is presented in Figure 4. This version had sections for youth opportunities and youth supports — reflecting both the importance of this language to the field and our need for generic constructs in which to cluster items to help data collectors make sense of the new instrument. However, it had no organizational sections other then several items describing the setting of the youth program — physical space, health, and safety. (It's curious to note that these were eventually moved back into the youth opportunities sections.) This July 2002 version of the youth instrument was used for data collection in a pilot study during 2002–2003.

During the summer of 2003, we developed the organizational sections of the youth instrument based on what we had already developed and by incorporating topics recommended to us by staff at Youth Development Strategies, Inc. Work during the summer of 2003 led to the version 2.79 of the Youth PQA. By the fall of 2003, we took bound copies of this version to the Search Institute conference in San Jose to show during our presentations there. In the fall of 2003, we began using version 2.79 of the youth instrument to collect data for the wave 1 sample. Figure 5 shows what the items of the Youth PQA Version 2.7 looked like at that time. Note the addition of the organizational items in sections three, four, and five and small changes in the Youth Opportunities and Youth Supports items.

Figure 5. Youth PQA Version 2.79 Item Tally

I. PROGRAM OFFERING: *Youth Opportunities*
Youth experience psychological & emotional safety.
Activities help develop a sense of belonging.
Opportunities to be actively engaged.
Opportunities to participate in small groups.
Opportunities to set goals and make plans.
Opportunities to make choices based on interests.
Opportunities for youth to reflect.
Opportunities for youth leadership.
Opportunities to practice communication skills

II. PROGRAM OFFERING: *Youth Supports*
Warm and caring atmosphere.
Physical space is appropriate for program offering.
Activities planned, presented, and paced for youth.
Support in learning and using new skills.
Variety of encouragement strategies.
Staff and youth as partners.
Staff as positive role models and mentors.
Youth-centered approach to resolve conflicts.
Staff support youth in defining and achieving success.

III. ORGANIZATION: *Policies & Structure*
Staff qualifications.
Staff orientation, meetings, and professional development.
Staffing patterns promote and sustain youth-staff relationships.
Schedules of program sessions are reliable and well publicized.
Program locations, schedules, and costs facilitate youth access.
Ongoing program improvement.
Communication and collaboration with parents, community, and schools.

IV. **ORGANIZATION**: *Activities*
Program offerings interest youth and build multiple skills.
Social norms promote sense of belonging and psychological safety.
High expectations.
Opportunities to influence setting and activities.
Opportunities to influence structure and policy.

V. ORGANIZATION: *Setting*
Safe and healthy physical environment for youth.
Emergency procedures and supplies
Physical environment accommodates program offerings.

We had been in contact with the Michigan Department of Education about the state's 21st Century afterschool projects. There was hope that we would be invited to use the Youth PQA in the state's 21st Century projects and that the Youth PQA would serve as a program quality monitoring tool for that program.

In January of 2004, the Michigan Department of Education set up a conference session in Lansing for High/Scope to train the state education consultants and contract evaluators in the Youth PQA. Using Youth PQA 2.79, we went to Lansing to train about a dozen consultants from the Michigan department of education and several representatives from the MSU evaluation team, as well as a handful of representatives from the state's 21st Century programs. This was a difficult and critically important moment in the tool's development. Version 2.79 quickly showed itself to have many ambiguities and other problems. The audience, while supportive, pointed out contradictions in our system of interpreting Youth PQA items and scores. We came back from the Michigan Department of Education training knowing that we still had a long way to go in making the items of the Youth PQA clear enough for people to use in assessing youth programs.

In late February and early March we held review sessions with some of the Youth PQA data collectors, most of whom were High/Scope certified trainers and expert practitioners. This was quite an important experience because it was the first time that we felt were getting direct feedback from the data collectors who had been using the Youth PQA in the field. For many items and descriptors, they pointed out areas of confusion, inconsistency, and simple clarity problems.

Figure 6. YPQA Version 2.7

I-A. Youth experience psychological & emotional safety in program activities.

☐ Activities with competitive elements result in predominately negative expressions (e.g., bragging, teasing, "junk-talking," non-verbal disrespect).	☐ Activities with competitive elements result in both positive and negative expressions (or none at all).	☐ Activities with competitive elements result in predominately positive expressions (e.g., mutual respect across competitors, encouraging statements, teamwork, camaraderie, everyone has a chance to participate).
☐ There is a sense of emotional hostility among members of the group (e.g., youth talk or behave negatively toward others).	☐ There is not a sense of emotional hostility, but there is not a clear sense of mutual respect and support.	☐ There is a clear sense of mutual respect and support.
☐ There is a sense of emotional hostility (e.g., criticisms and slurs) that has overtones of religious, ethnic, class, gender, or sexual orientation bias.	☐ There are indications of bias against others of different religions, ethnicity, class, gender, and/or sexual orientation but not hostility.	☐ There is a clear sense of mutual respect and support for others of different religions, ethnicity, class, gender, or sexual orientation.

YPQA Version 2.8

I-A. Psychological and emotional safety are promoted in program activities.

☐ The emotional **climate** of the session(s) is predominantly **negative** (disrespectful, tense, exclusive, even angry or hostile with negative behaviors such as, rudeness, bragging, cutting down, "trash talking," negative gestures or pushing that are not mediated by either youth or staff).	☐ The emotional climate of the session(s) is neutral or a balance of both positive and negative.	☐ The emotional **climate** of the session(s) is predominantly **positive** (e.g., mutually respectful, relaxed, supportive, with teamwork, camaraderie, inclusiveness, and an absence of negative behaviors). Any playful negative behaviors are mediated by staff or youth.
☐ There is explicit evidence (e.g., criticisms and slurs) of religious, ethnic, class, gender, or sexual orientation bias.	☐ There is implicit evidence (cliques, avoidance of others) of religious, ethnic, class, gender, or sexual orientation bias.	☐ There is no evidence of **bias** but rather mutual respect and support for and inclusion of others of different religions, ethnicity, class, gender, or sexual orientation.

In February, we also had a staff member conduct a detailed analysis of inconsistently scored items. He examined numerous completed Youth PQAs from wave 1 data collection, and analyzed the relationship between the evidence they cited and the scores they had given for particular items. He was able to identify a number of particularly problematic items.

In March of 2004, we began a major rewrite based on our feedback from the Michigan Department of Education consultants, from the data collectors meeting, and from the analysis of consistency. Figure 6 shows a sample rubric and how it changed during this rewrite. This rewrite of the Youth PQA would become the version for the wave 2 data collection that was to begin in the summer of 2004. The structure of this version (2.83) is shown in Figure 7.

We also had initial results from the wave 1 study available regarding item performance. In particular, we noted that that health and safety items, including psychological safety, were largely independent of the other items under youth opportunities and supports. This was demonstrated by factor loadings that were almost entirely different from the factor loadings of the other support and opportunity items.

Figure 7. YPQA Version 2.83 Item Tally

I. PROGRAM OFFERING: *Youth Opportunities*
A. Promoting psychological and emotional safety
B. Helping to develop a sense of belonging
C. Providing opportunities to be actively engaged
D. Providing opportunities to participate in small groups
E. Providing opportunities to set goals and make plans
F. Providing opportunities to make choices based on interests
G. Providing opportunities to reflect
H. Opportunities for developing leadership skills

II. PROGRAM OFFERING: *Youth Supports*
A. Providing a warm and caring atmosphere
B. Physical space is appropriate for program offering
C. Activities are planned, presented, and paced for youth
D. Supporting youth in developing skills
E. Encouraging youth through supportive strategies
F. Engaging youth as partners
G. Using youth-centered approaches to resolving conflicts

III. ORGANIZATION: *Policies & Structure*
A. Staff qualifications support a positive youth development focus
B. Staff orientation, meetings, and professional development
C. Staffing patterns promote and sustain youth-staff relationships
D. Schedules of program sessions are reliable and well publicized
E. Program locations, schedules, and costs facilitate youth access
F. Commitment to ongoing program improvement
G. Communicating and collaborating with parents, community, and schools

IV. ORGANIZATION: *Activities*
A. Program offerings tap youth interests to build multiple skills
B. Social norms promote sense of belonging and psychological safety
C. Supporting youth in meeting high expectations
D. Opportunities for youth to influence setting and activities
E. Opportunities for youth to influence structure and policy

V. ORGANIZATION: *Setting*
A. Safe and healthy physical environment for youth
B. Emergency procedures and supplies
C. Physical environment accommodates program offerings

By July 2004, the wave 2 data collection was off to a bumpy start. We were having difficulty scheduling observations in some of the summer programs because many of them were apt to change or be canceled at the last minute. We also faced the logistical task of collecting data for the organizational sections of the Youth PQA instrument.

We decided that during wave 2 data collection, the field observers would use only the first two sections of the Youth PQA. These sections comprise the items under youth opportunities and youth supports. In order to avoid repeating the interviews — since we were observing more than one offering in many organizations — these would be done separately. We had to gear up to conduct these interviews. Staff worked during July to prepare the interview items and the interview protocols. This meant repairing them and changing some of the organizational items that were giving data collectors difficulty.

Earlier, we prepared, rather hastily, an interview protocol for data collectors to use while conducting the interviews for the organizational items. This protocol was really just a set of discussion starters that would begin a discussion of the topic in question with the program director and lead indirectly

to the items that were to be rated for the organizational sections. This early protocol had many weaknesses. Most notably, many of the questions came across as leading to a preferred answer. As a result, it appeared that interviewees could guess the "right answer" and tell the interviewers what they wanted to hear rather than what was actually going on.

Staff worked to create an improved interview protocol. For budgetary feasibility, we aimed to conduct the interviews by telephone. The interview protocol involved a preliminary introduction and outline that would be faxed to the program director so that he or she would know what kinds of questions to prepare for what kinds of information to have ready. Staff would then set up a date and time for the interview and call.

By August of 2004, we'd gotten additional feedback from the data collectors on version 2.83 that was being used for the validation study. As a quality control measure, we had two staff members review each Youth PQA after it was scored and provide feedback to the data collector before they went back for their next visit to the youth program. Through this process, we reviewed 22 completed Youth PQAs, over 60 staff hours. This direct experience with the quality of the data and the observed difficulties and inconsistencies provided a rich source of feedback that could further the development of the Youth PQA items.

Figure 8. YPQA Version 3.2 Item Tally

I. PROGRAM OFFERING: *Youth Opportunities*
A. Promoting psychological and emotional safety
B. Helping to develop a sense of belonging
C. Providing opportunities to be actively engaged
D. Providing opportunities to participate in small groups
E. Providing opportunities to set goals and make plans
F. Providing opportunities to make choices based on interests
G. Providing opportunities to reflect
H. Opportunities for developing leadership skills

II. PROGRAM OFFERING: *Youth Supports*
A. Providing a warm and caring atmosphere
B. Activities are planned, presented, and paced for youth
C. Supporting youth in developing skills
D. Encouraging youth through supportive strategies
E. Engaging youth as partners
F. Using youth-centered approaches to resolving conflicts

III. PROGRAM OFFERING: *Safety*
A. Safe and healthy physical environment for youth
B. Emergency procedures and supplies
C. Physical environment accommodates program offerings

IV. ORGANIZATION: *Policies & Structure*
A. Staff qualifications support a positive youth development focus
B. Staff orientation, meetings, and professional development
C. Staffing patterns promote and sustain youth-staff relationships
D. Schedules of program sessions are reliable and well publicized
E. Program locations, schedules, and costs facilitate youth access
F. Commitment to ongoing program improvement
G. Communicating and collaborating with parents, community, and schools

V. ORGANIZATION: *Activities*
A. Program offerings tap youth interests to build multiple skills
B. Social norms promote sense of belonging and psychological safety
C. Supporting youth in meeting high expectations
D. Opportunities for youth to influence setting and activities
E. Opportunities for youth to influence structure and policy

In August of 2004, we met for one more major rewrite of the Youth PQA items. Even though the validation study was already under way, we felt that we had information about Youth PQA items that needed to be incorporated. Because we would be getting quite a lot of data in the fall of 2004 we could afford to make changes in the Youth PQA even at this late date. For this meeting, staff members had carefully noted which items of the Youth PQA were producing clarity or consistency issues. We went through the instrument, item by item. We took these comments and suggestions and revised once more, the items of the Youth PQA, producing version 3.2 (see Figure 8).

Besides many language improvements, the main change in 3.2 is the clarification and grouping of healthy and safety items. This subscale remained in later versions. Version 3.2 was used for the wave 2 data collection.

Early examinations of the data for waves 1 and 2 led to a complete restructuring of the Youth PQA subscales. This was a return to our some of our original ideas but with a new empirical theory of increasing depth of quality (i.e., safety is basic and common and gets high scores in most programs; engagement is the pinnacle of positive practice, is uncommon, and gets low scores in most programs) and had strongly supported by item clustering data. We produced version 4.0 and began to share the new construct with our stakeholders. Figure 9 shows the structure of 4.2 (overall structure from 4.0 to 4.2 changed very little).

Figure 9. YPQA Version 4.2 Item Tally

FORM A

I. Safe Environment
A. Psychological and emotional safety is promoted.
B. The physical environment is safe and free of health hazards.
C. Appropriate emergency procedures and supplies are present.
D. Program space and furniture accommodate the activities offered.
E. Healthy food and drinks are provided.

II. Supportive Environment
F. Staff provide a welcoming atmosphere.
G. Session flow is planned, presented, and paced for youth.
H. Activities support active engagement.
I. Staff support youth in building new skills.
J. Staff support youth with encouragement.
K. Staff use youth-centered approaches to reframe conflict.

III. Interaction
L. Youth have opportunities to develop a sense of belonging.
M. Youth have opportunities to participate in small groups.
N. Youth have opportunities to act as group facilitators and mentors.
O. Youth have opportunities to partner with adults.

IV. Engagement
P. Youth have opportunities to set goals and make plans.
Q. Youth have opportunities to make choices based on their interests.
R. Youth have opportunities to reflect.

FORM B

V. Youth-Centered Policies and Practices
A. Staff qualifications support a positive youth development focus.
B. Program offerings tap youth interests and build multiple skills.
C. Youth have an influence on the setting and activities in the organization.
D. Youth have an influence on the structure and policy of the organization.

VI. High Expectations for Youth and Staff
E. Organization promotes staff development.
F. Organization promotes supportive social norms.
G. Organization promotes high expectations for young people.
H. Organization is committed to ongoing program improvement.

VII. Access
I. Staff availability and longevity with the organization support youth-staff relationships.
J. Schedules are in effect.
K. Barriers to participation are addressed.
L. Organization communicates with families, other organizations, and schools.

The most significant occurrence involving version 4.0 happened in Washington, D.C., on May 19, 2005. At a meeting convened by the Forum for Youth Investment and supported by the W. T. Grant Foundation, key stakeholders met to discuss quality, and examined early findings of the Youth PQA study. The instruments were very well received, and we began the process of developing the story around what we've learned from the project.

Based on a final item-by-item examination by our instrument development team, and feedback received, we refined the final instrument. After going through an editing and publication process, the final printed version will be 4.2 (see Figure 9).

However, because of our ongoing work with the Michigan Department of Education 21st Century Community Learning Centers program, we became convinced that one of the primary constituencies for the this type of instrument would be afterschool programs serving elementary aged children. For this reason, a next round of revisions was undertaken between January and August 2005 to revise several items for use with younger children so that an observational version of the Youth PQA is available for use with children in grades K–4. The "younger youth (grade K–4)" version is structured to parallel the original Youth PQA version 4.2 with four subscales of the same name and about 80% identical content.