Measuring Youth Program Quality

A Guide to Assessment Tools, Second Edition

EXECUTIVE SUMMARY

Nicole Yohalem and Alicia Wilson-Ahlstrom, The Forum for Youth Investment
with Sean Fischer, New York University
and Marybeth Shinn, Vanderbilt University

the forum
FOR YOUTH INVESTMENT

Published by The Forum for Youth Investment
January 2009

About the Forum for Youth Investment

The Forum for Youth Investment is a nonprofit, nonpartisan “action tank” dedicated to helping communities and the nation make sure all young people are Ready by 21® – ready for college, work and life. Informed by rigorous research and practical experience, the Forum forges innovative ideas, strategies and partnerships to strengthen solutions for young people and those who care about them. A trusted resource for policy makers, advocates, researchers and practitioners, the Forum provides youth and adult leaders with the information, connections and tools they need to create greater opportunities and outcomes for young people.

The Forum was founded in 1998 by Karen Pittman and Merita Irby, two of the country’s top leaders on youth issues and youth policy. The Forum’s 25-person staff is headquartered in Washington D.C. in the historic Cady-Lee House with a satellite office in Michigan and staff in Missouri, New Mexico, Virginia and Washington.

Measuring Youth Program Quality

A Guide to Assessment Tools, Second Edition

Nicole Yohalem and Alicia Wilson-Ahlstrom, The Forum for Youth Investment
with Sean Fischer, New York University
and Marybeth Shinn, Vanderbilt University

Suggested Citation:

Yohalem, N. and Wilson-Ahlstrom, A. with Fischer, S. and Shinn, M. (2009, January). *Measuring Youth Program Quality: A Guide to Assessment Tools, Second Edition.* Washington, D.C.: The Forum for Youth Investment.

Please contact the Forum for Youth Investment at The Cady-Lee House, 7064 Eastern Ave, NW, Washington, D.C. 20012-2031, Phone: 202.207.3333, Fax: 202.207.3329, Web: www.forumfyi.org, Email: youth@forumfyi.org for information about reprinting this publication and information about other publications.

Acknowledgements

The authors would like to thank the following project advisors who helped develop the original scope of work, provided input into the interview protocol and report outline and reviewed the original report in draft form:

- Carol Behrer, Iowa Collaboration for Youth Development
- Priscilla Little, Harvard Family Research Project
- Elaine Johnson, National Training Institute for Community Youth Work
- Jeffrey Buehler, Missouri Afterschool State Network
- Bob Pianta, University of Virginia
- Marybeth Shinn, Vanderbilt University

We are especially grateful for the contributions of Sean Fischer and Marybeth Shinn, who took the lead on reviewing the technical properties of each instrument and drafted the technical sections of both the first (2007) and second editions of this report.

The developers of each of the tools described in this report also deserve thanks, for their willingness to share their materials, talk with us and review drafts. Thanks to Julie Goldsmith, Amy Arbreton, Beth Miller, Wendy Surr, Ellen Gannett, Judy Nee, Peter Howe, Suzanne Goldstein, Ajay Khashu, Liz Reisner, Ellen Pechman, Christina Russell, Rhe McLaughlin, Sara Mello, Thelma Harms, Charles Smith, Deborah Vandell and Kim Pierce.

Thanks to Karen Pittman for her guidance and suggestions throughout the project and to several Forum staff members, including Nalini Ravindranath and Laura Mattis for their assistance in the layout, design and editing process.

Finally, thanks to the William T. Grant Foundation for supporting this work and in particular to Bob Granger, Vivian Tseng and Ed Seidman, whose ideas, suggestions and encouragement were critical in transforming this from an idea to a final product.

Table of Contents

Introduction

With the after-school and youth development fields expanding and maturing over the past several years, program quality assessment has emerged as a central theme. This interest in program quality is shared by practitioners, policy makers and researchers in the youth-serving sector.

From a research perspective, more evaluations are including an assessment of program quality and many have incorporated setting-level measures (where the object of measurement is the program, not the participants) in their designs. At the policy level, decision-makers are looking for ways to ensure that resources are allocated to programs likely to have an impact and are increasingly building quality assessment and improvement expectations into requests for proposals and program regulations. At the practice level, programs, organizations and systems are looking for tools that help concretize what effective practice looks like and allow practitioners to assess, reflect on and improve their programs.

With this growing interest in program quality has come an increase in the number of tools available to help programs and systems assess and improve quality. Given the size and diversity of the youth-serving sector, it is unrealistic to expect that any one quality assessment tool will fit all programs or circumstances. While diversity in available resources is positive and reflects the evolution of the field, it also makes it important that potential users have access to good information to help guide their decision-making.

Over the last several years, we at the Forum have found ourselves regularly fielding questions related to program quality assessment including what tools exist, what it takes to use them and what might work best under what conditions. The need to offer guidance to the field in terms of available resources has become increasingly clear.

This guide was designed to compare the purpose, structure, content and technical properties of several youth program quality assessment tools. It builds on work we began in this area five years ago, as well as recent

The following tools are included in the guide at this time:

Assessing Afterschool Program Practices Tool (APT)
National Institute on Out-of-School Time and Massachusetts Department of Elementary & Secondary Education

Communities Organizing Resources to Advance Learning Observation Tool (CORAL)
Public/Private Ventures

Out-of-School Time Observation Tool (OST)
Policy Studies Associates, Inc.

Program Observation Tool (POT)
National AfterSchool Association

Program Quality Observation Scale (PQO)
Deborah Lowe Vandell and Kim Pierce

Program Quality Self-Assessment Tool (QSA)
New York State Afterschool Network

Promising Practices Rating Scale (PPRS)
Wisconsin Center for Education Research and Policy Studies Associates, Inc.

Quality Assurance System® (QAS)
Foundations, Inc.

School-Age Care Environment Rating Scale (SACERS)
Frank Porter Graham Child Development Institute and Concordia University, Montreal

Youth Program Quality Assessment (YPQA)
David P. Weikart Center for Youth Program Quality

work conducted by the Harvard Family Research Project to document and compile quality standards for middle school programs[1].

Criteria for Inclusion

With any compendium comes the challenge of determining what to include. Our first caveat is that we plan to continue revising this guide over time, in part because in its current form it is not inclusive of the universe of relevant tools and in part because a great deal of innovation is currently underway. Many of the tools included in the review will be revised or will undergo further field testing in the next 1-2 years.

Our criteria for inclusion in the guide were as follows:

- ***Tools that are or that include setting-level observational measures of quality.*** We are particularly interested in direct program observation as a means for gathering specific data about program quality and in particular, staff practice. Therefore this review does not feature other methodological approaches to measuring quality (e.g., surveying participants, staff or parents about the program).
- ***Tools which are applicable in a range of school and community-based program settings.*** We did not include tools that are designed to measure how well a specific model is being implemented (sometimes referred to as fidelity) or have limited applicability beyond specific organizations or approaches.
- ***Tools that include a focus on social processes within programs.*** Many of the tools in this guide address some static regulatory or licensing issues (e.g., policies related to staffing, health and safety). However, we are particularly interested in tools that address social processes or the interactions between and among people in the program.
- ***Tools which are research-based.*** All of the tools included are "research-based" in the sense that their development was informed by relevant child/youth development literature. Although we are particularly interested in instruments with established technical properties (e.g., reliability, validity), not all of those included fit this more rigorous definition of "research-based."

Purpose and Contents of the Guide

We hope this compendium will provide useful guidance to practitioners, policy makers, researchers and evaluators in the field as to what options are available and what issues to consider when selecting and using a quality assessment tool. It focuses on the purpose and history, content, structure and methodology, technical properties and user considerations for each of the instruments included, as well as a brief description of how they are being used in the field. For each tool, we aim to address the following key questions:

Purpose and History. Why was the instrument developed – for whom and in what context? Is its primary purpose program improvement? Accreditation? Evaluation? For what kinds of programs, serving what age groups, is it appropriate for?

Content. What kinds of things are measured by the tool? Is the primary focus on the activity, program or organization level? What components of the settings are emphasized – social processes, program resources, or the arrangement of those resources (Seidman, Tseng & Weisner, 2006)? How does it align with the National Research Council's positive developmental settings framework[2] (2002)?

Structure and Methodology. How is the tool organized and how do you use it? How are data collected and by whom? How do the rating scales work and how are ratings determined? Can the tool be used to generate an overall program quality score?

Technical Properties. Is there any evidence that different observers interpret questions in similar ways (reliability)? Is there any evidence that the tool measures what it is supposed to measure (validity)?

[1] Westmoreland, H. & Little, P. (2006). *Exploring quality standards for middle school after school programs: What we know and what we need to Know: A summit report.* Harvard Family Research Project; Cambridge, MA. Retrieved online at www.gse.harvard.edu/hfrp/content/projects/afterschool/conference/summit-2005-summary.pdf.

[2] National Research Council and Institute of Medicine. (2002). Community programs to promote youth development. Eccles, J. and Gootman, J., eds. Washington, DC: National Academy Press.

See the Appendix for a "psychometrics dictionary" that defines relevant terminology and explains why technical properties are an important consideration.

User Considerations. How easy is the tool to access and use? Does it come with instructions that are understandable for practitioners as well as researchers? Is training available on the instrument itself or on the content covered by it? Are data collection, management and reporting services available? What costs are associated with using the tool?

In the Field. How is the tool being applied in specific programs or systems?

To ensure that the guide is useful to a range of audiences with different purposes and priorities, we have provided both in-depth and summary level information in a variety of formats.

For each tool, we provide both a one page "at-a-glance" summary as well as a longer description. The at-a-glance summaries or longer tool descriptions can stand alone as individual resources. Should you decide to use one of these instruments or want to take a closer look at two or three, you could pull these sections out and share with key stakeholders.

We also provide cross-instrument comparison charts and tables for those who want to get a sense of what the landscape of program quality assessment tools looks like. The Cross-Cutting Observations section that follows compares the instruments across most of the categories listed above (purpose, content, structure, technical properties, user considerations). While definitions of quality do not differ dramatically across the instruments, there are notable differences in some of these other areas which we try to capture.

Updated Content

In this edition of the guide, we update the summaries of nine assessment tools featured in the original March 2007 edition, and add an additional tool – the Communities Organizing Resources to Advance Learning (CORAL) Observation Tool) – developed by Public/Private Ventures. This edition also includes refined definitions of validity and a discussion regarding some of the limitations of traditional methods of establishing reliability.

Since our original publication, there has been a flurry of activity related to the development and use of the various tools. Almost all of the tool developers have continued to work on either technical or practical aspects of their assessment tools, as well as on related resources to support practitioner use of these tools.

These changes demonstrate continued investment on the part of developers in making tools more accessible and user-friendly to programs and systems trying to implement quality assessment and improvement. Changes that have been made or are in development since 2007 include:

- Further psychometric testing of the reliability and validity of measures (OST; YPQA)
- Development and/or expansion of resources to support the use of various tools (APT; POT; QSA; QAS)
- Development and/or expansion of the availability of web-based tools and resources (QAS; QSA; YPQA)
- Aligning quality assessment tools with other measures to create a package of compatible tools (APT)
- Restructuring of the framework and/or scales (APT; OST)
- Expanding access by translating a tool into different languages (SACERS)
- Development of brother/sister tools targeting different age groups (YPQA; SACERS)

We hope this compendium will provide useful guidance to practitioners, policymakers, researchers and evaluators in the field as to what options are available and what issues to consider when selecting and using a quality assessment tool. We look forward to updating the compendium again as this work advances.

Cross-Cutting Comparisons

Although the individual tool descriptions include what we hope is useful information about several different program quality assessment instruments, their level of detail may be daunting, particularly without a sense of the broader landscape of resources. Some of the individualized information about each tool can be further distilled in ways that may help readers understand both the broader context of program quality assessment and where individual tools fall within that context. We were not able to collect completely comparable information about all instruments in every topic area, but in those cases where we were, we have summarized and compared that information in narrative and charts.

TOOL DEVELOPERS KEY

APT: Assessing Afterschool Program Practices Tool
National Institute on Out-of-School Time and Massachusetts Department of Elementary & Secondary Education

CORAL: Communities Organizing Resources to Advance Learning Observation Tool
Public/Private Ventures

OST: Out-of-School Time Observation Tool
Policy Studies Associates, Inc.

POT: Program Observation Tool
National AfterSchool Association

PQO: Program Quality Observation Scale
Deborah Lowe Vandell and Kim Pierce

QSA: Program Quality Self-Assessment Tool
New York State Afterschool Network

PPRS: Promising Practices Rating Scale
Wisconsin Center for Education Research and Policy Studies Associates, Inc.

QAS: Quality Assurance System®
Foundations, Inc.

SACERS: School-Age Care Environment Rating Scale
Frank Porter Graham Child Development Institute and Concordia University, Montreal

YPQA: Youth Program Quality Assessment
David P. Weikart Center for Youth Program Quality

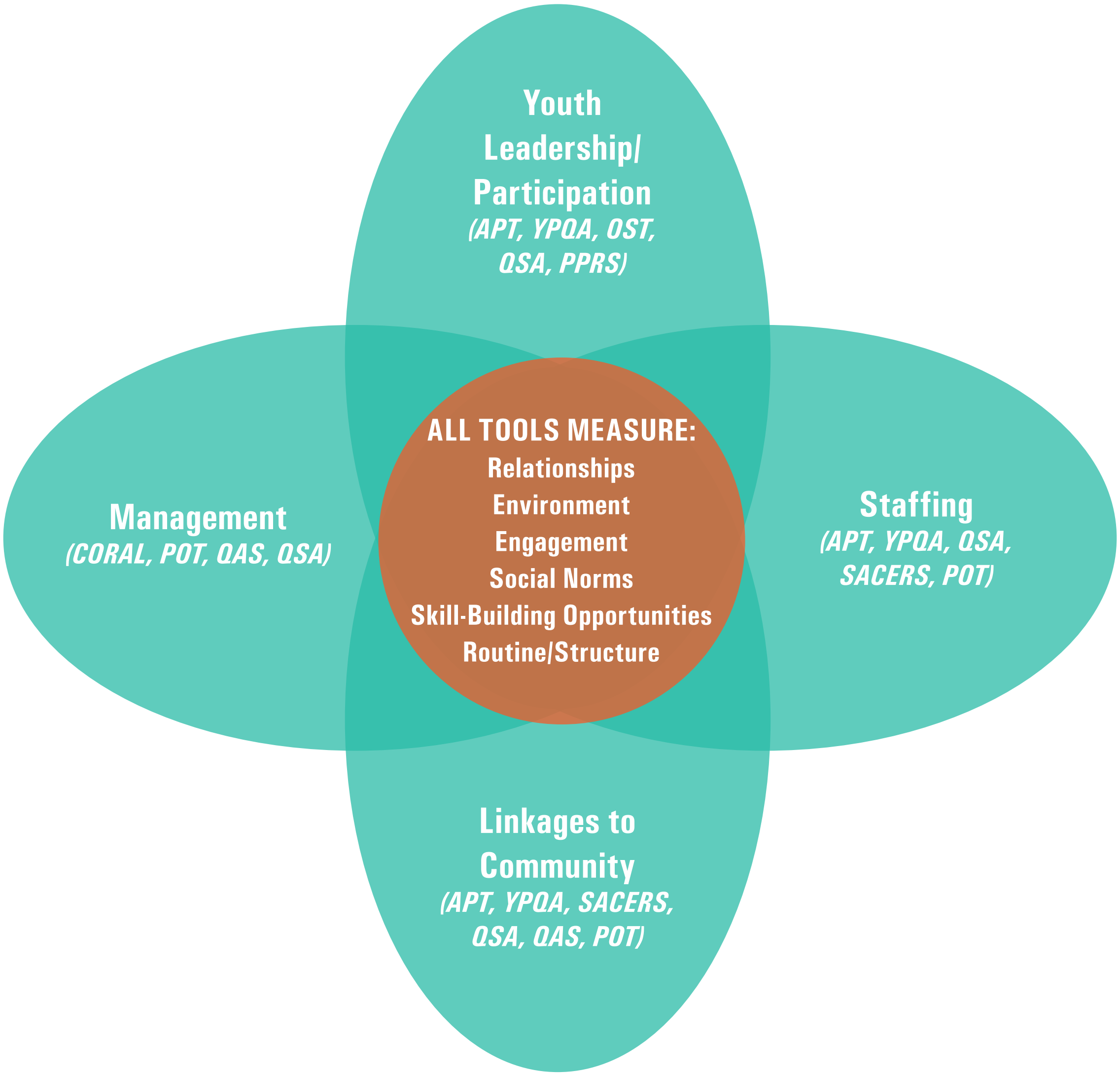
Figure 1: Target Age and Purpose

Most of the tools included in this review were developed primarily for self-assessment and program improvement purposes. Some, however, were developed with program monitoring or accreditation as a key goal and several were developed exclusively for use in research. Many have their roots in early childhood assessment (SACERS, POT, PQO) while others draw more heavily on youth development and/or education literature (APT, CORAL, OST, PPRS, QAS, QSA, YPQA). While the majority of tools were designed to assess programs serving a broad range of children (often K–12 or K–8), some are tailored for more specific age ranges.

	Program Target Age	Primary Purpose		
	Grades Served	*Improvement*	*Monitoring/ Accreditation*	*Research/ Evaluation*
Assessing Afterschool Program Practices Tool (APT)	Grades K–8	✓	✓	
Communities Organizing Resources to Advance Learning Observation Tool (CORAL)	Grades K–5		✓	✓
Out-of-School Time Observation Tool (OST)	Grades K–12			✓
Program Observation Tool (POT)	Grades K–8	✓	✓	
Program Quality Observation Scale (PQO)	Grades 1–5			✓
Program Quality Self-Assessment Tool (QSA)	Grades K–12	✓		
Promising Practices Rating Scale (PPRS)	Grades K–8			✓
Quality Assurance System (QAS)	Grades K–12	✓		
School-Age Care Environment Rating Scale (SACERS)	Grades K–6	✓	✓	✓
Youth Program Quality Assessment (YPQA)	Grades 4–12	✓	✓	✓

Figure 2: Common and Unique Content

There is reasonable consensus across instruments about the core features of settings that matter for development. All of the tools included in this review measure six core constructs (at varying levels of depth): relationships, environment, engagement, social norms, skill building opportunities and routine/structure. The content of most of the instruments aligns well with the National Research Council's features of positive development settings framework[3] (2002), which has helped contribute to the growing consensus around elements of quality that has emerged since then. In terms of what components of settings the tools emphasize (Seidman et al, 2006), all include a focus on social processes. Although only a subset emphasize program resources, several include items related to the arrangement of resources within the setting.


Youth Leadership/ Participation
(APT, YPQA, OST, QSA, PPRS)
Management
(CORAL, POT, QAS, QSA)
ALL TOOLS MEASURE:
Relationships
Environment
Engagement
Social Norms
Skill-Building Opportunities
Routine/Structure
Staffing
(APT, YPQA, QSA, SACERS, POT)
Linkages to Community
(APT, YPQA, SACERS, QSA, QAS, POT)

[3] This report included a list of "features of positive developmental settings" culled from frequently cited literature. It has contributed to the emerging consensus about the components of program quality.

Figure 3: Methodology

Many of the tools included in this review follow a similar structure. They tend to be organized around a core set of topics or constructs, each of which is divided into several items, which are then described by a handful of more detailed indicators. Some variation does exist, however. For example, the PQO includes a unique time sampling component.[2] While most tools are organized around features of quality, some are not. For example, while the APT addresses a core set of quality features, the tool itself is organized around the program's daily routine (e.g., arrival, transitions, pick-up). Observation is the primary data collection method for each of the instruments in this review, although several rely upon interview, questionnaire or document review as additional data sources.

	Target Users		Data Collection Methods			
	Program Staff	*External Observers*	*Observation*	*Interview*	*Questionnaire*	*Document Review*
Assessing Afterschool Program Practices Tool (APT)	✓	✓	✓		✓	
Communities Organizing Resources to Advance Learning Observation Tool (CORAL)		✓	✓			
Out-of-School Time Observation Tool (OST)		✓	✓			
Program Observation Tool (POT)	✓	✓	✓		✓	✓
Program Quality Observation Scale (PQO)		✓	✓			
Program Quality Self-Assessment Tool (QSA)	✓		✓			✓
Promising Practices Rating Scale (PPRS)		✓	✓			
Quality Assurance System (QAS)	✓	✓	✓	✓		✓
School-Age Care Environment Rating Scale (SACERS)	✓	✓	✓	✓		
Youth Program Quality Assessment (YPQA)	✓	✓	✓	✓		

[2] The time sampling method has observers go through a cycle of selecting individual participants (ideally at random) to observe for brief periods of time and document their experiences.

Figure 4: Strength of Technical Properties

Most of the instruments have some information showing that if different observers watch the same program practices, they will score the instrument similarly (internal consistency and interrater reliability). Few, however, have looked at other aspects of reliability that are of interest when assessing the strength of a program quality measure. Several of the instruments have promising findings to consider in terms of validity – meaning they have made some effort to demonstrate that the instrument accurately measures what it is supposed to measure. See the accompanying glossary on page 15 and the Appendix for more detailed definitions of psychometric terms.

	Score Distributions	*Interrater Reliability*	*Test-Retest Reliability*	*Internal Consistency**	*Convergent Validity*	*Concurrent/ Predictive Validity*	*Validity of Scale Structure**
Assessing Afterschool Program Practices Tool (APT)		✓✓†				✓✓†	
Communities Organizing Resources to Advance Learning Observation Tool (CORAL)	✓✓✓			✓✓		✓✓✓	✓✓
Out-of-School Time Observation Tool (OST)	✓✓✓	✓✓✓		✓✓✓		✓✓	✓✓
Program Observation Tool (POT)		✓✓✓†	✓✓✓†	✓✓✓†	✓✓†		
Program Quality Observation Scale (PQO)	✓✓✓	✓✓✓	✓✓	✓✓✓	✓✓✓	✓✓	N/A
Program Quality Self-Assessment Tool (QSA)							
Promising Practices Rating Scale (PPRS)	✓✓✓	✓✓		✓✓✓		✓✓	N/A
Quality Assurance System (QAS)							
School-Age Care Environment Rating Scale (SACERS)		✓✓✓		✓✓✓	✓✓	✓✓	
Youth Program Quality Assessment (YPQA)	✓✓✓	✓✓	✓✓✓	✓✓	✓✓✓	✓✓	✓✓✓

* This type of evidence is only relevant for instruments with a lot of items that would be useful if organized into scales.
† Psychometric information is not based on the instrument in its current form, so its generalizability may be limited.

Key

= No Evidence

✓✓✓ = Evidence of this property is strong by general standards

✓✓ = Evidence of this property is moderate by general standards, promising but limited or mixed (strong on some items or scale, weaker on others)

✓ = Evidence of this property is weaker than desired

Figure 5: Technical Glossary

	What is it?	Why is it Useful?
Score Distributions	The dispersion or spread of scores from multiple assessments for a specific item or scale.	In order for items and scales (sets of items) to be useful, they should be able to distinguish difference between programs. If almost every program scores low on a particular scale, it may be that the items make it "too difficult" to obtain a high score and, as a result, don't distinguish between programs on this dimension very well.
Interrater Reliability	How much assessments by different trained raters agree when observing the same program at the same time.	It is important to use instruments that yield reliable information regardless of the whims or personalities of individual observers If findings depend largely on who is rating the program (rater A is more likely to give favorable scores than rater B), it is hard to get a sense of the program's actual strengths and weaknesses.
Test-Retest Reliability	The stability of an instrument's assessments of the same program over time.	If an instrument has strong test-retest reliability than the score it generates should be stable over time. This is important because we want changes in scores to reflect real changes in program quality. The goal is to avoid situations where an instrument is either too sensitive to subtle changes that may hold little significance, or insensitive to important long-term changes.
Internal Consistency	The cohesiveness of items forming an instrument's scales.	Scales are sets of items within an instrument that jointly measure a particular concept. If, however, the items within a given scale are actually conceptually unrelated to each other, then the overall score for that scale may not be meaningful.
Convergent Validity	The extent to which an instrument compares favorably with another instrument (preferably one with demonstrated validity strengths) measuring identical or highly similar concepts.	It is important to use an instrument that generates accurate information about what you are trying to measure. If two instruments are presumed to measure identical or highly similar concepts, we would expect programs that receive high scores on one measure to also receive high scores on the other.
Concurrent/ Predictive Validity	The extent to which an instrument is related to distinct theoretically important concepts and outcomes in expected ways.	If an instrument accurately measures high program quality then one can expect it to predict better outcomes for the youth participating in the program. The instruments findings should also be related to distinct, theoretically important variables and concepts in expected ways.
Validity of Scale Structure	The extent to which items statistically group together in expected ways to form scales.	It is helpful to know exactly which concepts an instrument is measuring. Factor analysis can help determine if one scale actually incorporates more than one related concept or if different items can be combined because they are essentially measuring the same thing.

Figure 6: Training and Support for Users

Six of the ten instruments included in this review are free to users and available to download from the Internet; the other four have various costs associated with their use. In most, but not all cases, training is available (at a fee) for those interested in using the tool. Many come with user-friendly manuals that explain how to use the instrument; in some cases these materials are still under development. In several cases, the developers of the tools also provide data collection, management and reporting services at additional cost to users. Details about such considerations are included in the individual tool descriptions.

	Cost	*Training Available*	*Estimated Time Necessary to Train Overservers to Generate Reliable Scores*	*Estimated Minimum Observation Time Needed to Generate Sound Data*	*Data Collection, Management and Reporting Available*
Assessing Afterschool Program Practices Tool (APT)	Free*	Yes	4 hour training plus 2 program observations	1 afternoon (2-3 hours)	No
Communities Organizing Resources to Advance Learning Observation Tool (CORAL)	Free	No	2 days	3-4 hours	No
Out-of-School Time Observation Tool (OST)	Free	No †	8-18 hours, depending on experience	3 hours	No †
Program Observation Tool (POT)	$300 Advancing Quality Kit	Yes	2.5-3 days	3-5 hours (for self-assessment)	No
Program Quality Observation Scale (PQO)	Free	No †	2 hours plus 2-4 observations & 2-4 time samples, depending on experience	1.5 hours observation & .5 hours time sampling	No †
Program Quality Self-Assessment Tool (QSA)	Free	Yes	2 hours ††	N/A	No
Promising Practices Rating Scale (PPRS)	Free	No †	2 hours plus 2-4 observations, depending on experience	2 hours	No †
Quality Assurance System (QAS)	$75 Annual Site License	Yes	2-3 hours ††	1 afternoon (2-3 hours)	Yes
School-Age Care Environment Rating Scale (SACERS)	$15.95 SACERS Booklet	Yes	4-5 days	3 hours	Yes
Youth Program Quality Assessment (YPQA)	$39.95 YPQA Starter Pack	Yes	2 days	4 hours	Yes

* A fee structure may be developed over time, once additional materials are completed.

† Training and data services have only been made available in the context of specific research projects.

†† These are estimates of time necessary to prepare observers; developers of these tools have not trained "to reliability."

At-a-Glance Summaries

Detailed descriptions of the ten assessment tools are provided in the next section. Here we offer one-page summaries to copy and share. Each summary follows a common format.

Assessing Afterschool Program Practices Tool (APT)
National Institute on Out-of-School Time and Massachusetts Department of Elementary & Secondary Education

Communities Organizing Resources to Advance Learning Observation Tool (CORAL)
Public/Private Ventures

Out-of-School Time Observation Tool (OST)
Policy Studies Associates, Inc.

Program Observation Tool (POT)
National AfterSchool Association

Program Quality Observation Scale (PQO)
Deborah Lowe Vandell and Kim Pierce

Program Quality Self-Assessment Tool (QSA)
New York State Afterschool Network

Promising Practices Rating Scale (PPRS)
Wisconsin Center for Education Research and Policy Studies Associates, Inc.

Quality Assurance System® (QAS)
Foundations, Inc.

School-Age Care Environment Rating Scale (SACERS)
Frank Porter Graham Child Development Institute and Concordia University, Montreal

Youth Program Quality Assessment (YPQA)
David P. Weikart Center for Youth Program Quality

Assessing Afterschool Program Practices Tool

Developed by NIOST and the Massachusetts Department of Elementary & Secondary Education

Overview:

The Assessment of Afterschool Program Practices Tool (APT) is designed to help practitioners examine and improve what they do in their program to support young people's learning and development. It examines those program practices that research suggests relate to youth outcomes (e.g., behavior, initiative, social relationships). A research version of the APT (the APT-R) was developed in 2003-2004. This more user-friendly self-assessment version was developed in 2005.

Primary Purpose(s):

Program Improvement; Monitoring/Accreditation

Program Target Age:

Grades K–8

Relevant Settings:

Both structured and unstructured programs that serve elementary and middle school students during the non-school hours.

Content:

The APT measures a set of 15 program-level features and practices that can be summarized into five broad categories – program climate, relationships, approaches and programming, partnerships and youth participation.

Structure:

The 15 program features addressed by the APT are measured by two tools – the observation instrument (APT-O) and questionnaire (APT-Q). The APT-O guides observations of the program in action, while the APT-Q examines aspects of quality that are not easily observed and guides staff reflection on those aspects of practice and organizational policy.

Methodology:

Items that are observable within a given program session (typically one full afternoon) are assessed in the APT-O. The APT-Q is a questionnaire to gather information about planning, frequency and regularity of program offerings and opportunities and frequency of connections with families and school. Both the APT-O and APT-Q have four-point scales, though flexibility is encouraged for users who find the scales not useful for their purposes. Depending on what part of the tool(s) is being used, the scales measure how characteristic an item is of the program, the consistency of an item or the frequency of an item. For each item, concrete descriptors illustrate what a score of 1, 2, 3 or 4 looks like.

Technical Properties:

While no psychometric information is available for the current self-assessment version of the APT, some is available on the research version (APT-R) on which it is based. For the APT-R, interrater reliability was moderate and preliminary evidence of concurrent and predictive validity is available. NIOST has plans for further testing of the APT.

User Considerations:

Ease of Use

- "Cheat sheets" demonstrate link between quality and outcomes.
- Instrument is extremely flexible in terms of administration, use of scales, number of observations, etc.
- The instrument is designed for users to make observations in just one program session.
- The instrument can be used as part of a package including an outcomes tool and data tracking system.

Available Supports

- Training on both the APT itself and the youth development principles embedded in the instrument is available through NIOST.
- Packaging and pricing information about training on the instrument is available from NOIST for organizations not already affiliated with the APT.

For More Information:

www.niost.org/content/view/1572/282/
or www.doe.mass.edu/21cclc/ta

Communities Organizing Resources to Advance Learning Observation Tool

Developed by Public/Private Ventures

Overview:

The CORAL observation tool was designed by Public/Private Ventures (P/PV) for the CORAL after-school initiative funded by the James Irvine Foundation. The tool was developed for research purposes and was primarily used in a series of evaluation studies on the CORAL after-school initiative. The primary purpose of the observations was to monitor fidelity to the Balanced Literacy Model and change in quality and outcomes over time. The tool was used in two ways: 1) observation of literacy instruction and 2) observation of programming in support of literacy. Though the CORAL observation tool was designed to help observers measure the impact of after-school programs on academic achievement, it has applications for observing quality in a wide variety of settings.

Primary Purpose:

Research/Evaluation

Program Target Age:

Grades K–5

Relevant Settings:

Structured literacy-based programs, both school and community-based.

Content:

The CORAL observation tool documents the connection between the quality of the program, fidelity to the Balanced Literacy Model and the academic outcomes of participants.

Structure:

The CORAL observation tool is structured around five key constructs of quality – adult-youth relations, effective instruction, peer cooperation, behavior management and literacy instruction. The tool is divided into five parts. The first three – the activity description form, characteristics form and the activity checkbox form – are focused on describing the activity as well as participant and staff behavior. The second two components include an activity scale and an overall assessment form, and are completed after a 90-minute observation period.

Methodology:

Each construct is based on a five-point rating scale. The activity description form, characteristics form and activity checkbox form are filled out before an activity is observed, and contain the most informative aspects of the activity. The activity scale and overall assessment form are completed after a 90-minute observation session.

Technical Properties:

Evidence for score distributions and predictive validity is strong by general standards, and evidence for internal consistency and the validity of scale structure is promising but limited.

User Considerations:

Ease of Use

- Contains detailed instructions for conducting observations.
- Includes space for open-ended narratives.
- Scoring takes 3-4 hours, including completing the rating scales, related narratives and the overall assessment.

Available Supports

- Currently, training is limited to individuals involved in specific evaluations that employ the instrument.
- Public/Private Venture's website features a free download of materials in their Afterschool Toolkit.

For More Information:

www.ppv.org/ppv/initiative.
asp?section _ id=0&initiative _ id=29

Out-of-School Time Program Observation Tool

Developed by Policy Studies Associates, Inc.

Overview:

The Out-of-School Time Program Observation Tool (OST) was developed in conjunction with several research projects related to out-of-school time programming, with the goal of collecting consistent and objective data about the quality of activities through observation. Its design is based on several assumptions about high-quality programs – first that certain structural and institutional features support the implementation of high-quality programs and second that instructional activities with certain characteristics – varied content, mastery-oriented instruction and positive relationships – promote positive youth outcomes.

Primary Purpose:

Research/Evaluation

Program Target Age:

Grades K–12

Relevant Settings:

Varied school- and community-based after-school programs.

Content:

The OST documents and rates the quality of the following major components of after-school activities: interactions between youth and adults and among youth, staff teaching processes and activity content and structures.

Structure:

The first section of OST allows for detailed documentation of activity type, number and demographics of participants, space used, learning skills targeted, type of staff and the environmental context. The remainder of the tool assesses the quality of activities along five key domains including relationships, youth participation, staff skill building and mastery strategies and activity content and structure.

Methodology:

The OST observation instrument uses a seven-point scale to assess the extent to which each indicator is or is not present during an observation. Qualitative documentation, recorded on site, supplements the rating scales. Activity and quality indicator data from the OST observation instrument is used in conjunction with related survey measures.

Technical Properties:

Evidence for interrater reliability is strong by general standards, as is evidence for score distributions and internal consistency. Evidence for concurrent validity and the validity of the scale structure is promising but limited.

User Considerations:

Ease of Use

- Free and available online.
- Tool includes an introduction and basic procedures for use.
- Includes some technical language but has been used by both researchers and practitioners.
- Raters must observe approximately 3 hours of programming to generate sound data.
- Observers can be trained to generate reliable observations through 8-16 hours of training, depending on level of experience.

Available Supports

- Training is limited to individuals involved in specific evaluations that employ the instrument.
- Additional non-observational measures related to after-school programming are available from PSA that can be used in conjunction with the OST.

For More Information:

www.policystudies.com/studies/youth/OST%20
Instrument.html

Program Observation Tool

Developed by the National AfterSchool Association

Overview:

The Program Observation Tool is the centerpiece of the National AfterSchool Association's (NAA) program improvement and accreditation process and is designed specifically to help programs assess progress against the Standards for Quality School-Age Care. Developed in 1991 by NAA and the National Institute on Out-of-School Time, the tool was revised and piloted before the accreditation system began in 1998.

Primary Purpose(s):

Program Improvement; Monitoring/Accreditation

Program Target Age:

Grades K–8

Relevant Settings:

School and center-based after-school programs.

Content:

The Program Observation Tool measures 36 "keys of quality," organized into six categories. Five are assessed primarily through observation: human relationships; indoor environment; outdoor environment; activities; and safety, health and nutrition. The sixth – administration – is assessed through questionnaire/interview. The tool reflects NAA's commitment to holistic child development and its accreditation orientation.

Structure:

The five quality categories that are the focus of the tool are measured using one instrument that includes the 20 relevant keys and a total of 80 indicators (four per key). If a program is going through the accreditation process, the administration items are assessed separately, through questionnaire/interview.

Methodology:

The rating scale captures whether each indicator is true all of the time, most of the time, sometimes or not at all. Specific descriptions of what a 0, 1, 2 or 3 looks like are not provided, but descriptive statements help clarify the meaning of each indicator. Programs seeking accreditation must assign an overall program rating based on individual scores and guidelines are provided for observers to reconcile and combine scores. For accreditation purposes, the program/activities and safety/nutrition categories are "weighted."

Technical Properties:

No psychometric evidence is available on the POT itself, but there is information about the ASQ (Assessing School-Age Childcare Quality), from which the POT was derived. Overall, evidence for interrater and test-retest reliability is strong by general standards. Following revisions to the scales, evidence for internal consistency was also strong. Preliminary evidence of concurrent validity is also available for the ASQ.

User Considerations:

Ease of Use

- Accessible language and format developed with input from practitioners.
- When used for self-assessment, observation and scoring takes roughly 3-5 hours.
- A self-study manual provides detailed guidance on instrument administration.
- The package costs approximately $300 (additional costs for full accreditation).

Available Supports

- The POT is part of an integrated set of resources for self-study and accreditation.
- The full accreditation package provides detailed guides, videos and other supports.
- Beginning in September 2008, accreditation is offered through the Council on Accreditation.
- NAA currently offers training that covers the Program Observation Tool through its day-long Endorser Training (NAA recommends two and a half days of training in order to ensure reliability).
- Some NAA state affiliates offer training for programs interested in self-assessment and improvement.

For More Information:

http://naaweb.yourmembership.com/?page=NAAAccreditation

Program Quality Observation Scale

Developed by Deborah Lowe Vandell & Kim Pierce

Overview:

The Program Quality Observation Scale (PQO) was designed to help observers characterize the overall quality of an after-school program environment and to document individual children's experiences within programs. The PQO has been used in a series of research studies and has its roots in Vandell's observational work in early child care settings.

Primary Purpose:

Research/Evaluation

Program Target Age:

Grades 1–5

Relevant Settings:

Varied school- and community-based after-school programs.

Content:

The PQO focuses primarily on social processes and in particular, three components of quality of children's experiences inside programs: relationships with staff, relationships with peers and opportunities for engagement in activities.

Structure:

The tool has two components – qualitative ratings focused on the program environment and staff behavior (referred to as "caregiver style") and time samples of children's activities and interactions. While program environment ratings are made of the program as a whole, caregiver style ratings are made separately for each staff member observed.

Methodology:

All items are all assessed through observation (although the PQO has always been used in tandem with other measures that rely on different kinds of data). Program environment and caregiver style ratings are made using a four-point scale and users are given descriptions of what constitutes a 1, 2, 3 or 4 for three aspects of environment and four aspects of caregiver style. In the time sample of activities, activity type is recorded using 19 different categories and interactions are assessed and coded along several dimensions.

Technical Properties:

Evidence for interrater reliability, score distributions, internal consistency and convergent validity is strong by general standards and evidence for test-retest reliability and concurrent/predictive validity is promising but mixed.

User Considerations:

Ease of Use

- Free and available for use.
- The PQO was developed with a research audience in mind. Manual includes basic instructions for conducting observations and completing forms but has not been tailored for general or practitioner use at this time.
- Qualitative ratings of environment and staff require a minimum of 90 minutes observation time. Completing the time samples as outlined takes a minimum of 30 minutes for an experienced observer.

Available Supports

- Training has only been made available in the context of a specific research study.
- Data collection, management or reporting have only been available in the context of a specific study.
- The authors have developed a range of related measures that can be used in conjunction with the PQO (e.g., physical environment questionnaire; staff, student and parent surveys).

For More Information:

http://childcare.gse.uci.edu/des4.html

Program Quality Self-Assessment Tool

Developed by the New York State Afterschool Network

Overview:

The Program Quality Self-Assessment Tool (QSA) was developed exclusively for self-assessment purposes (use for external assessment and formal evaluation purposes is discouraged). The QSA is intended to be used as the focal point of a collective self-assessment process that involves all program staff. Soon after it was created in 2005, the state of New York began requiring that all 21st CCLC-funded programs use it twice a year for self-assessment purposes.

Primary Purpose:

Program Improvement

Program Target Age:

Grades K–12

Relevant Settings:

The full range of school and community-based after-school programs. The QSA is particularly relevant for programs that intend to provide a broad range of services as opposed to those with either a very narrow focus or no particular focus (e.g., drop-in centers).

Content:

The QSA is organized into 10 essential elements of effective after-school programs, including environment/climate; administration/organization; programming/activities; and youth participation/ engagement, among others. A list of standards describes each element in greater detail. The elements represent a mix of activity-level, program-level and organizational-level concerns.

Structure:

Each of the QSA's 10 essential elements is further defined by a summary statement which is then followed by between 7 and 18 quality indicators. The four-point rating scale used in the QSA is designed to capture performance levels for each indicator. Indicators are also considered standards of practice, so the goal is to determine whether the program does or does not meet each of the standards.

Methodology:

While most essential elements are assessed through observation, the more organizationally focused elements such as administration, measuring outcomes/evaluation and program sustainability/growth are assessed primarily through document review. Users are not encouraged to combine scores for each element to determine a global rating, because the tool is intended for self-assessment only.

Technical Properties:

Beyond establishing face validity, the instrument's psychometric properties have not been researched.

User Considerations:

Ease of Use

- Practitioners led the development of the QSA; language and format are clear and user-friendly.
- The tool is free and downloadable and includes an overview and instructions.
- The tool is scheduled for a revision which will target length and guidance on determining ratings.

Additional Supports

- The New York State Afterschool Network has developed a user guide, which provides a self-guided walk-through of the tool.
- Programs can contact the New York State Afterschool Network to receive referrals for technical assistance in using the instrument.
- Programs are encouraged to use the QSA in concert with other formal or informal evaluative efforts.
- NYSAN trainings are organized around the 10 elements featured in the instrument, so practitioners can easily find professional development opportunities that connect to the findings in their self-assessment.

For More Information:

www.nysan.org

Promising Practices Rating Scale

Developed by the Wisconsin Center for Education Research & Policy Studies Associates, Inc.

Overview:

The Promising Practices Rating Scale (PPRS) was developed in the context of a study of the relationship between participation in high quality after-school programs and child and youth outcomes. The tool was designed to help researchers document type of activity, extent to which promising practices are implemented within activities and overall program quality. The PPRS builds directly on earlier work by Deborah Lowe Vandell and draws upon several other observation instruments included in this report.

Primary Purpose:

Research/Evaluation

Program Target Age:

Grades K–8

Relevant Settings:

Varied school- and community-based after-school programs.

Content:

The PPRS focuses primarily on social processes occurring at the program level (other tools in the PP assessment system are available to collect other kinds of information). The tool addresses activity type, implementation of promising practices and overall program quality. The practices at the core of the instrument include supportive relations with adults, supportive relations with peers, level of engagement, opportunities for cognitive growth, appropriate structure, over-control, chaos and mastery orientation.

Structure:

The first part of the instrument focuses on activity context. Observers code things like activity type, space, skills targeted, number of staff and youth involved. Observers then add a brief narrative description of the activity. The core of the PPRS is where observers document to what extent certain exemplars of promising practice are present in the program.

Methodology:

All items in the scale are addressed through observation, with an emphasis first on activities and then more broadly on the implementation of promising practices by staff within the program. Each area of practice is divided into specific exemplars (positive and negative) with detailed indicators. Ratings are assigned at the overall practice level using a four-point scale. Observers then review their ratings of promising practices across multiple activities and assign an overall rating for each practice area and the overall program.

Technical Properties:

Strong evidence for score distribution and internal consistency of the average overall score has been established. Promising but limited evidence of moderate interrelater reliability and predictive validity have also been established.

User Considerations:

Ease of Use

- Free and available for use.
- The PPRS was developed with a research audience in mind. Manual includes basic instructions for conducting observations and completing forms but has not been tailored for general or practitioner use at this time.
- In the study the PPRS was developed for, formal observation time totaled approximately two hours per site, with additional hours spent reviewing notes and assigning ratings.

Available Supports

- Training has only been made available in the context of a specific study.
- Data collection, management or reporting has only been available in the context of a specific study.
- The authors have developed a range of related measures that can be used in conjunction with the PPRS (e.g., physical environment questionnaire; staff, student and parent surveys).

For More Information:

http://childcare.gse.uci.edu/des3.html

Quality Assurance System®

Developed by Foundations, Inc.

Overview:

The Quality Assurance System® (QAS) was developed to help programs conduct quality assessment and continuous improvement planning. Based on seven "building blocks" that are considered relevant for any after-school program, this Web-based tool is expandable and has been customized for particular organizations based on their particular focus. The QAS focuses on quality at the "site" level and addresses a range of aspects of quality from interactions to program policies and leadership.

Primary Purpose:

Program Improvement

Program Target Age:

Grades K–12

Relevant Settings:

A range of school- and community-based programs.

Content:

The various components of quality that the QAS focuses on are considered "building blocks." The seven core building blocks include: program planning and improvement; leadership; facility and program space; health and safety; staffing; family and community connections; and social climate. Three additional "program focus building blocks" that reflect particular foci within programs are also available.

Structure:

The QAS is divided into two parts. Part one – program basics – includes the seven core building blocks.
For each, users are given a brief description of the importance of that aspect of quality and then the building block is further subdivided into between five and eight elements, each of which gets rated. Part two of the tool – program focus – consists of the three additional building blocks and its structure parallels that of part one. Ratings for the QAS are made using a four-point scale from unsatisfactory (1) to outstanding (4).

Methodology:

Filling out the QAS requires a combination of observation, interview and document review. Users follow a five-step process for conducting a site visit and collecting data, which includes observation of the program in action and a review of relevant documents. Once ratings for each element are entered into the computer, scores are generated for each building block – rather than a single score for the overall program – reflecting the tool's emphasis on identifying specific areas for improvement.

Technical Properties:

Beyond establishing face validity, research about the instrument's psychometric properties has not been conducted.

User Considerations:

Ease of Use

- The QAS is flexible and customizable, with built-in user-friendly features.
- The instruction guide walks the user through basic steps for using the system.
- The $75 annual licensing fee covers two assessments and cumulative reports.
- Multi-site programs can generate site comparison reports.

Available Supports

- Foundations, Inc. offers online sessions and in-person trainings.
- Once a QAS site license is purchased, programs can receive light phone technical assistance free of charge from staff.
- Programs that wish to have their assessment conducted by trained assessors can purchase this service under contract with Foundations, Inc.
- The QAS is available in a Web-based format allowing users to enter data and immediately generate basic graphs and analyses.

For More Information:

http://qas.foundationsinc.org/start.asp?st=1

School-Age Care Environment Rating Scale

Developed by Frank Porter Graham Child Development Institute & Concordia University, Montreal

Overview:

The School-Age Care Environment Rating Scale (SACERS), published in 1996 and updated periodically, is one of a series of quality rating scales developed by researchers at the Frank Porter Graham Child Development Institute. SACERS focuses on "process quality" or social interactions within the setting, as well as features related to space, schedule and materials that support those interactions. The SACERS can be used by program staff as well as trained external observers or researchers.

Primary Purpose(s):

Program Improvement; Monitoring/Accreditation; Research/Evaluation

Program Target Age:

Grades K–8

Relevant Settings:

A range of program environments including child care centers, school-based after-school programs and community-based organizations.

Content:

SACERS is based on the notion that quality programs address three "basic needs": protection of health and safety, positive relationships and opportunities for stimulation and learning. The seven sub-scales of the instrument include space and furnishings; health and safety; activities; interactions; program structure; staff development; and a special needs supplement.

Structure:

The SACERS scale includes 49 items, organized into seven subscales. All 49 items are rated on a seven-point scale, from "inadequate" to "excellent." Concrete descriptions of what each item looks like at different levels are provided. All of the sub-scales and items are organized into one booklet that includes directions for use and scoring sheets.

Methodology:

While observation is the main form of data collection, several items are not likely to be observed during program visits. Raters are encouraged to ask questions of a director or staff person in order to rate these and are provided with sample questions. For many items, clarifying notes help the user understand what they should be looking for. Observers enter scores on a summary score sheet, which encourages users to compile ratings and create an overall program quality score.

Technical Properties:

Evidence for interrater reliability and internal consistency is strong by general standards. Convergent and concurrent validity evidence is limited but promising.

User Considerations:

Ease of Use

- Accessible format and language.
- Includes full instructions for use, clarifying notes and a training guide.
- The cost of SACERS booklet is $15.95.
- Suggested time needed: three hours to observe a program and complete form.
- Guidance is offered on how to sample, observe and score to reflect multiple activities within a program.

Available Supports

- Additional score sheets can be purchased in packages of 30.
- Three and five-day trainings on SACERS structure, rationale and scoring.
- Guidance on how to conduct your own training is provided in booklet.
- Training to reliability takes 4-5 days, with reliability checks throughout.
- Access to a listserv through the Franklin Porter Graham Institute Web site.
- Large scale users can now use commercial software to enter/score data.
- With Web-based reporting system, individual assessments can be routed to a supervisor for quality assurance and feedback and aggregate analyses and reporting can be provided.

For More Information:

www.fpg.unc.edu/~ecers/

Youth Program Quality Assessment

Developed by the David P. Weikart Center for Youth Program Quality[4]

Overview:

The Youth Program Quality Assessment (YPQA) was developed by the High/Scope Educational Research Foundation and has its roots in a long lineage of quality measurement rubrics for pre-school elementary and now youth programs. The overall purpose of the YPQA is to encourage individuals, programs and systems to focus on the quality of the experiences young people have in programs and the corresponding training needs of staff. While some structural and organizational management issues are included in the instrument, the YPQA is primarily focused on what the developers refer to as the "point of service" – the delivery of key developmental experiences and young people's access to those experiences.

Primary Purpose(s):

Program Improvement; Monitoring/Accreditation; Research/Evaluation

Program Target Age:

Grades 4–12

Relevant Settings:

Structured programs in a range of school- and community-based settings.

Content:

Because of the focus on the "point of service," the YPQA emphasizes social processes – or interactions between people within the program. The majority of items are aimed at helping users observe and assess interactions between and among youth and adults, the extent to which young people are engaged in the program and the nature of that engagement. However the YPQA also addresses program resources (human, material) and the organization or arrangement of those resources within the program.

Structure:

The YPQA assesses seven domains using two overall scales. Topics covered include engagement, interaction, supportive environment, safe environment, high expectations, youth-centered policies and practices and access.

Methodology:

Items at the program offering level are assessed through observation. Organization level items are assessed through a combination of guided interview and survey methods.

The scale used throughout is intended to capture whether none of something (1), some of something (3) or all of something (5) exists. For each indicator, concrete descriptors illustrate what a score of 1, 3 or 5 looks like.

Technical Properties:

Evidence for score distributions, test-retest reliability, convergent validity and validity of scale structure is strong. Evidence for interrater reliability is mixed and evidence is promising but limited in terms of internal consistency and concurrent validity.

User Considerations:

Ease of Use

- Language and format of the tool are accessible.
- Administration manual with definitions of terms and scoring guidelines.
- The tool can be ordered online.
- Raters must observe for roughly four hours to generate sound data.
- Observers can be trained to generate reliable observations in two days.

Available Supports

- One-day basic and two-day intermediate YPQA training are available, with additional technical assistance available upon request.
- Youth development training that is aligned with tool content is available.
- Online "scores reporter" and a Web-based data management system are available.

For More Information:

www.highscope.org/content.asp?contentid=117

[4] The Weikart Center is a joint venture between the High/Scope Educational Research Foundation and the Forum for Youth Investment.

Appendix

Psychometrics: What are they and why are they useful?

By Sean Fischer

The youth program Janice works for is interested in self-assessment and is looking for a tool that measures the overall quality of the program. After looking over several options, she settles on an instrument that seems easy to use, with questions that seem relevant to the organization's goals. Unfortunately, she encounters a number of problems once she starts using the instrument. First, the observers interpret questions very differently, leading to disputes over their assessments of quality. Second, the individual item scores don't seem to form a coherent picture of the program. Third, the findings are unrelated to youth outcomes that should be directly related to program quality. All of these issues make Janice question whether the instrument measures program quality as well as it should.

The instrument Janice chose looked useful on the surface, but its field performance was not particularly helpful. Psychometric information might have helped Janice understand the strengths and weaknesses of the instrument before she used it. Psychometrics are statistics that help researchers evaluate instruments' field performances. Psychometric information can be divided into several categories.

Reliability

An instrument's ability to generate consistent answers or responses.

The most common analogy used to understand reliability is a game of darts. If a player's darts consistently land on the same location on the board, we would say that the dart player has excellent reliability (whether or not that place is the center of the board). The same is true for research instruments that yield predictable and consistent information. There are various types of reliability discussed below.

Interrater Reliability

The extent to which trained raters agree when evaluating the same program at the same time.

For accurate program assessments, users should choose instruments that yield reliable information regardless of the whims or personalities of individual raters. When findings depend largely on who is rating the program (e.g., if Rater A is more likely to give favorable scores than Rater B), it is hard to get a sense of the program's actual strengths and weaknesses. For this reason, organizations should consider the interrater reliability of various measures even if only one rater will be rating the program. Poor interrater reliability often stems from ambiguous questions that leave a lot of room for individual interpretation and such ambiguity is not always immediately apparent from looking at the instrument.

Several methods exist to measure interrater reliability. Many of the instruments in this report give the percentage that raters agree for a given item (allowing a one-point difference to count as agreement). While this method is common, it is not as useful as other statistics. When available, we instead report two other statistics known as kappa and intraclass correlation. Values of kappa near or above .70 indicate high reliability and this value is often considered the benchmark for a strong, reliable instrument. Other researchers state that kappa values starting at .60 indicate substantial/strong agreement, whereas values ranging from .40 to .59 indicate moderate agreement. Similar guidelines do not yet exist for the intraclass correlation, but this report considers values close to or above .50 to indicate high reliability.

The reason that percentage agreement does not sufficiently represent reliability is that it does not account for those instances where raters agree simply by chance, whereas kappa scores and intraclass correlations do. In many cases, what looks like high interrater agreement may actually have a low kappa score or intraclass correlation coefficient. When kappa scores or intraclass correlations are not available for an instrument, we provide an estimate of kappa. Readers should know that the estimate is the best possible score based on the available information, though it is possible the actual kappa score is much lower (indicating worse reliability).

It is important to note that interrater reliability statistics assume that all raters have been adequately trained on the instrument. Some instruments' developers offer training for raters. If you cannot receive formal training on an instrument, it is still important to train raters yourself before conducting an evaluation. Organizations can hold meetings to review each question individually and discuss what criteria are necessary to assign a score of 1, 2 or 3, etc. If possible, raters should go through "test evaluations" to practice using the instrument with scenarios that could occur in the program (ideally through videos, but such scenarios could also be written if detailed enough). When disagreement occurs on individual questions, raters should discuss why they chose to rate the program the way they did and come to a consensus. Practice evaluations will help raters get "on the same page" and have a mutual understanding of what to look for.

Test-Retest Reliability

The stability of an instrument's assessments of the same program over time.

If several after-school programs are each assessed two times, one month apart, the respective scores at both assessments would differ very little if the instrument had strong test-retest reliability. The strength of an instrument's test-retest reliability depends on both the sensitivity of the instrument and how much the program changes over time. If instruments are too sensitive to subtle changes in a program, test-retest reliability will be low and scores may differ widely between assessments even though the subtle changes driving this difference may hold little practical significance. On the other extreme, instruments with extremely high test-retest reliability may be insensitive to important long-term changes. As is the case with interrater reliability, several methods to measure test-retest reliability exist including percentage agreement, kappa and intraclass correlations, with the latter two being preferred.

Very few of the instruments in this report have undergone testing for this type of reliability. Because the time span between assessments has been relatively short for these instruments, test-retest reliability should be high.

Internal Consistency

The cohesiveness of items forming the instrument's scales.

An item is a specific question or rating and a scale is a set of items within an instrument that jointly measure a particular concept. For example, an instrument might include 10 items that are supposed to measure the friendliness of program staff and users would average or sum the 10 scores to get an overall "friendliness score." Because items forming a scale jointly measure the same concept, we can expect that the scores for each item will be related to all of the other items. For example, say that three of our "friendliness" items include: (1) How much does the staff member smile at children? (2) How much does the staff member compliment children? (3) How much does the staff member criticize children in a harsh manner? If the scale had high internal consistency, the scores for each question would "make sense" compared to the others (e.g., if the first question received a high score, we would expect that the second would also receive a high score and the third would receive a low score). In a scale with low internal consistency the items' scores are unrelated to each other. Low internal consistency suggests the items may not fit together in a meaningful way and therefore the overall score (e.g., average friendliness) may not be meaningful either.

The analogy of the dartboard is useful when understanding internal consistency. Think about the individual items as the darts: the aim of the thrower is meaningless if the darts land haphazardly across the board. In the same way, an overall score like average friendliness is meaningless if the items' scores do not relate to each other. The statistic that determines internal consistency is called Cronbach's alpha. For a scale to have acceptable internal consistency, it should be near or over the conventional cutoff of 0.70. Whereas interrater and test-retest reliabilities are important information for all instruments, internal consistency is only relevant for instruments with scales.

The Weikart Center (YPQA developer), among others (MacKenzie, S., Podsakoff, P., & Jarvis, C., 2005), has noted that internal consistency is only appropriate when the items are reflective of a larger concept rather than formative. For a more in-depth discussion of this requirement, readers should refer to the section on Additional Technical Considerations, found on pages 16-17 of this report.

Variation in Quality Across Different Contexts

Program quality may not be entirely uniform across different staff, different activities, or even different days of the week or months of the year. Even when two observers can agree on the level of quality that they are observing when both are observing precisely the same activity at the same time, they might come up with different ratings if they observe a different activity at a different time. Some instruments may also be particularly sensitive to some types of variation. As the Weikart Center and others have noted (Raudenbush, S., Martinez, A., Bloom, H., Zhu, P., & Lin, F., 2008), evidence about the ways that scores on a particular instrument vary within a program is important so that users know how to account for this variation (e.g., if an instrument's scores depend on the activity, then it is important to assess a wide range of activities in the program). For a more in-depth discussion of these issues, readers should refer to the section Additional Technical Considerations, found on pages 16-17 of this report.

Validity[17]

An instrument's ability to measure what it is supposed to measure.

If an instrument is supposed to measure program quality, then it would be valid if it yielded accurate information on this topic. However researchers have devised several different methods for establishing validity. The most common analogy used to understand validity again is the game of darts. While reliability is about the player consistently throwing darts to the same location, validity relates to whether or not the player is hitting the bull's eye. The bull's eye is the topic an instrument is supposed to measure. While reliability is essential, it is also important to know if an instrument is valid (dart players that consistently miss the board entirely may be reliable – they may hit the same spot over and over – but they are sure to lose the game!).

Sometimes an instrument may look like it measures one concept when in fact it measures something rather different or nothing at all. For example, an instrument might claim to measure after-school program quality, but it would not be particularly valid if it focused solely on whether children liked the program and were having fun.

Validity can be tricky to assess because the concepts of interest (e.g., program quality) are often not tangible or concrete. Unlike the case of reliability, there is no specific number that tells us about validity. These methods each assess different types of relationships that together give us confidence that the instrument is measuring what we think it measures. Next, we describe the different subtypes of validity.

Face Validity

Individuals' opinions of an instrument's quality.

This is the weakest form of validity because it does not involve direct testing of the instrument and is based on appearance only. One example of face validity in a medical context concerns taking a temperature. Today we know to do this with a thermometer. But think back a couple hundred years. At that time, feeling a patient's forehead would have seemed a much more valid measure of temperature than sticking a glass tube filled with mercury into the patient's mouth. How hot a forehead feels is a face valid measure of temperature, but few people today consider this method alone to be adequate. Instead, doctors rely on thermometers because they have been scientifically proven to be more accurate. Similarly, researchers and practitioners should consider other forms of validity when available before choosing an instrument.

Convergent Validity

The extent to which an instrument compares favorably with another instrument (preferably one

[17] Researchers often refer to the type of validity discussed in this report as Construct Validity, because it addresses whether an instrument adequately measures a specific concept or construct. Although other forms of validity exist, they are not addressed in this report.

with demonstrated validity strengths) measuring identical or highly similar concepts.

If two instruments are presumed to measure the same or similar concepts, we would expect programs that receive high scores on one measure to also receive high scores on the other. For example, imagine researchers have developed a new instrument (Instrument A) that is supposed to measure staff behavior management techniques in after-school programs. To determine its validity, researchers might compare Instrument A to Instrument B, which is already known to accurately measure staff's discipline strategies in after-school programs. Assuming that Instrument A is a valid measurement, we can expect that when Instrument B finds that programs rarely use appropriate discipline strategies, Instrument A will find that the same programs utilize poor behavior management techniques (and vice versa). If this were not the case, we would conclude that Instrument A probably does not adequately measure behavior management.

Concurrent and Predictive Validity

The extent to which an instrument is related to distinct theoretically important concepts and outcomes in expected ways.

If an instrument measures the quality of homework assistance in after-school programs, then children who attend high quality programs should have higher rates of homework completion (or perhaps grades) than children who attend low quality programs (assuming there is no difference between the children before starting the programs). Usually, theory and prior research findings help researchers determine which outcomes are most appropriate to examine with each instrument. Validity evidence is strongest when differences in the outcomes are detected after the initial program observations have been conducted (known as Predictive Validity). For example, imagine that two after-school programs are designed to improve children's grades, and that children attending these programs had similar grades at the beginning of the school year. After conducting program observations, researchers determined that one program was of high quality and the other was of low quality. If children attending the high quality program had higher grades at the end of the school year compared to the children attending the low quality program, this makes us more confident that the instrument accurately detected quality differences between the two programs.

Sometimes observations and related concepts are measured in the same time period (known as Concurrent Validity), particularly when the related concepts are expected to change simultaneously. However researchers generally prefer to see the hypothesized cause (program quality) come before the effect. When both are measured at the same time, it is more likely that there may be another explanation for the relationship.

Although similar in some ways, concurrent and predictive validity are separate from convergent validity. Whereas convergent validity compares two instruments that measure identical or highly similar concepts, concurrent and predictive validity refer to relationships between distinct concepts that we expect to be strong based on theory and prior research.

Validity of Scale Structure

The extent to which items statistically group together in expected ways to form scales.

As already stated, scales are composed of several items that, when averaged or summed, create an overall score of a specific concept. Determining whether scales adequately measure the concepts they claim to measure can be difficult, though conducting a factor analysis is one helpful way to do so. Factor analysis verifies that items go together the ways the developers thought they would by examining which items are similar to each other and which are different.

For example, imagine an instrument with two scales: Staff Communication Style and Staff Patience. Next, imagine that whenever staff are rated as having a harsh communication style toward children, they are also always rated as having little patience with children. Because of their high similarity, we would say that we are actually measuring one concept, not two, and it would make more sense to have one overall score (perhaps renamed Staff Attitudes Toward Children).

Factor analysis can also help determine if one scale actually incorporates more than one related concept. Imagine that we have an instrument with a scale called Homework Assistance, but our factor analysis finds that we actually have two separate concepts. We might discover that some items relate to Tutoring on Specific Subject Matter whereas another set relates to Teaching Study Skills. The reason that the validity of scale structure is important is because we want to know exactly which concepts our instrument measures.

Score Distribution

The dispersion or spread of scores from multiple assessments for a specific item or scale, including features such as the average score, the range of observed values and their concentration around particular point(s).

In order for items and scales to be useful, they should be able to distinguish differences between programs on a range of qualities. To achieve this, scores should not be "bunched up" on any particular place on the scale. For example, imagine that a particular instrument has a scale called Positive Child Behavior and users must rate, from 1 to 5, how true statements like "Children never stop helping each other" and "Children thank staff at every opportunity" are for a large number of programs. If almost every program scored low for this particular scale, we might argue the items are making it "too difficult" to obtain a high score and do not meaningfully distinguish between programs on this dimension. One solution would be to revise the items to better reflect program differences. The two sample items above might be revised to say "Children help each other when needed" and "Children appreciate help from staff."

Several important statistics help researchers understand whether scores are bunching up on the ends, including the average score (sometimes called the mean) and how spread out the scores are. For example, a scale or item would not be very useful for distinguishing between programs if the average score across many different programs was a 4.8 out of a possible 5.0. In addition, a scale or item might have an average of 3.5, but it would be less useful if the scores only ranged between 3 and 4 instead of a larger spread between 1 and 5.

The Forum for Youth Investment
The Cady-Lee House
7064 Eastern Avenue N.W.
Washington, D.C. 20011
Phone. 202.207.3333
Fax. 202.207.3329
youth@forumfyi.org
www.forumfyi.org