



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

**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.

©2009 The Forum for Youth Investment. All rights reserved. Parts of this report may be quoted or used as long as the authors and the Forum for Youth Investment are recognized. No part of this publication may be reproduced or transmitted for commercial purposes without prior permission from 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	6
Updated Content	7
Cross-Cutting Comparisons	10
At-a-Glance Summaries	19
<i>Assessing Afterschool Program Practices Tool</i>	<i>20</i>
<i>Communities Organizing Resources to Advance Learning Observation Tool</i>	<i>21</i>
<i>Out-of-School Time Observation Instrument</i>	<i>22</i>
<i>Program Observation Tool</i>	<i>23</i>
<i>Program Quality Observation Tool</i>	<i>24</i>
<i>Program Quality Self-Assessment Tool</i>	<i>25</i>
<i>Promising Practices Rating Scale</i>	<i>26</i>
<i>Quality Assurance System</i> ®.....	<i>27</i>
<i>School-Age Care Environment Rating Scale</i>	<i>28</i>
<i>Youth Program Quality Assessment</i>	<i>29</i>
Individual Tool Descriptions	30
<i>Assessing Afterschool Program Practices Tool</i>	<i>31</i>
<i>Communities Organizing Resources to Advance Learning Observation Tool</i>	<i>38</i>
<i>Out-of-School Time Observation Instrument</i>	<i>43</i>
<i>Program Observation Tool</i>	<i>48</i>
<i>Program Quality Observation Tool</i>	<i>53</i>
<i>Program Quality Self-Assessment Tool</i>	<i>59</i>
<i>Promising Practices Rating Scale</i>	<i>63</i>
<i>Quality Assurance System</i> ®.....	<i>68</i>
<i>School-Age Care Environment Rating Scale</i>	<i>72</i>
<i>Youth Program Quality Assessment</i>	<i>77</i>
References	85
Appendix	87

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 work conducted by the Harvard Family Research Project

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

to document and compile quality standards for middle school programs (Westmoreland, H. & Little, P., 2006).

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 framework¹ (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)? See the Appendix for a “psychometrics dictionary” that defines relevant terminology and explains why technical properties are an important consideration.

¹ 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.

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.

Figure 1: Target Age and Purpose

Figure 2: Common and Unique Content

Figure 3: Methodology

Figure 4: Strength of Technical Properties

Additional Technical Considerations

Figure 5: Technical Glossary

Figure 6: Training and Support for Users

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

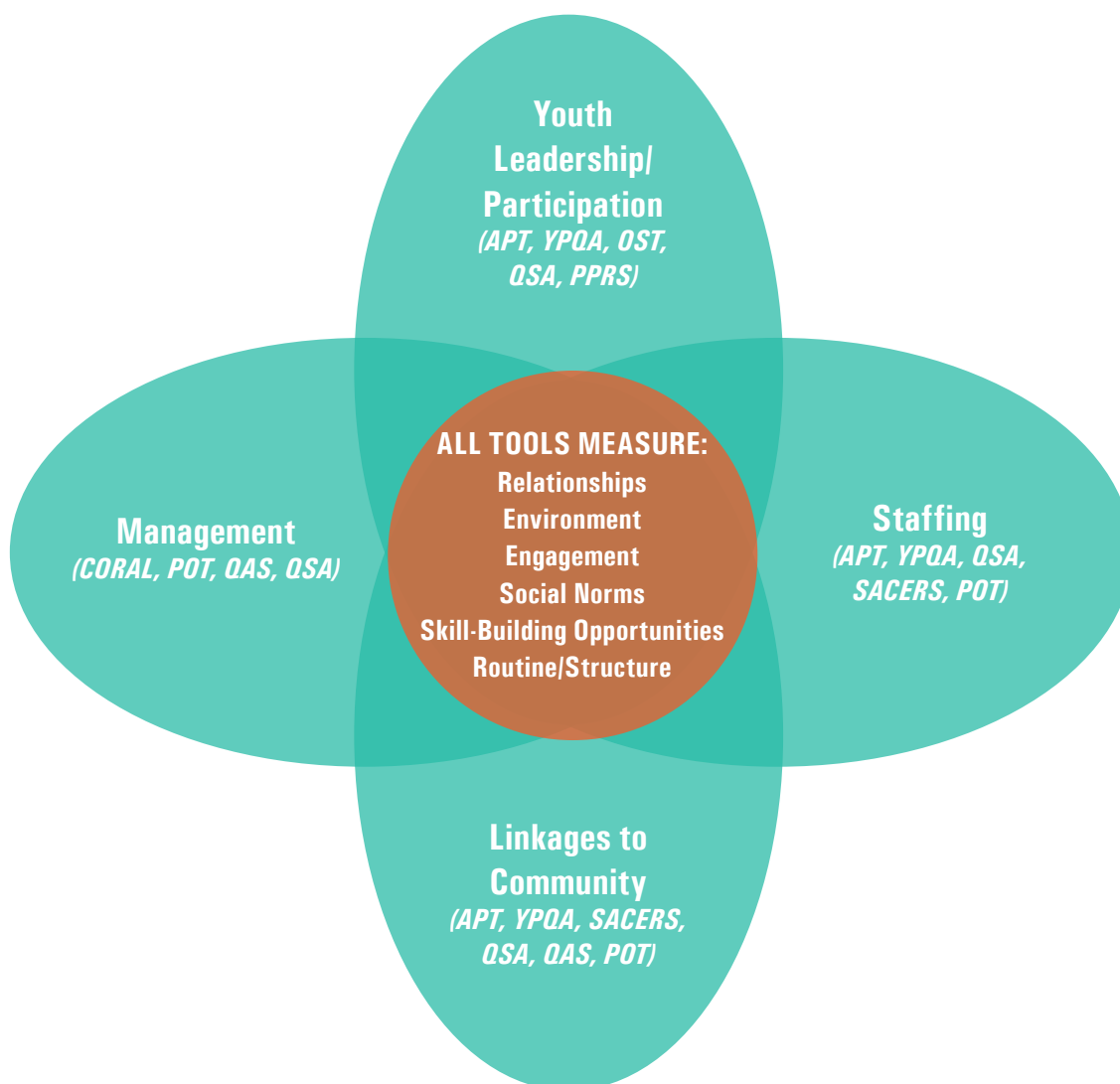


Figure 3: Methodology

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
<i>Assessing Afterschool Program Practices Tool (APT)</i>	✓	✓	✓		✓	
<i>Communities Organizing Resources to Advance Learning Observation Tool (CORAL)</i>		✓	✓			
<i>Out-of-School Time Observation Tool (OST)</i>		✓	✓			
<i>Program Observation Tool (POT)</i>	✓	✓	✓		✓	✓
<i>Program Quality Observation Scale (PQO)</i>		✓	✓			
<i>Program Quality Self-Assessment Tool (QSA)</i>	✓		✓			✓
<i>Promising Practices Rating Scale (PPRS)</i>		✓	✓			
<i>Quality Assurance System (QAS)</i>	✓	✓	✓	✓		✓
<i>School-Age Care Environment Rating Scale (SACERS)</i>	✓	✓	✓	✓		
<i>Youth Program Quality Assessment (YPQA)</i>	✓	✓	✓	✓		

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.

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

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

* 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.

Figure 5: Technical Glossary

	<i>What is it?</i>	<i>Why is it Useful?</i>
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.

Additional Technical Considerations

Many instruments in this report have strong reliability and validity evidence using traditionally accepted techniques. However, as with any field, new methods are often introduced to advance our understanding of reliability and validity. In this section, we discuss some of the limitations of traditional methods that have been highlighted by researchers at the David P. Weikart Center for Youth Program Quality³ (YPQA developer), what these methods can and cannot tell us about an instrument's reliability and validity, and how new methods are addressing these issues.

Variations in Quality Across Different Contexts

In an ideal world, scores we obtain from program quality instruments would always be perfectly accurate. Unfortunately, reality tends to be messier because many factors influence assessments. For example, different raters may not perceive observations in the same way and thus give different scores to the same questions. Different staff or different activities might get different scores, if we could observe them all, but typically we cannot do so. Another possible issue could be that program staff interact with children differently at the beginning of the year (when they do not know the children well yet) versus the end of the year. When these influences are unaccounted for when using an instrument, they are collectively known as "error variance". When an instrument is reliable, its scores are not influenced by much error.

Unfortunately, traditional reliability methods, including the ones used by instruments in this report, do not account for all possible sources of variation in scores (thereby increasing the inaccuracy of the instrument), as discussed by Steve Raudenbush and his colleagues (for a readable treatment of this topic, see Martinez & Raudenbush, 2008). Charles Smith at the Weikart Center has done preliminary work that examines sources of variation for the YPQA, including whether ratings are different during earlier program sessions versus later program sessions. He has also examined

whether it is enough to observe one type of activity within an agency versus observing a broad range of activities (readers who are interested in specific findings should refer to the technical summary of the YPQA, see pages 78-82).

When we know how an instrument is influenced by all these factors, we can take steps to reduce error. For example, if an instrument's scores vary widely depending on which activities we are observing, then we should observe a wide range of activities. If scores depend on the time of day, then we should conduct observations at multiple times throughout the day. By accounting for these additional influences on program quality, we reduce error and obtain more accurate scores. At this point in time, the YPQA is the only instrument to have preliminary information on external sources of variation beyond interrater reliability.

Understanding Assumptions about Internal Consistency

Several instruments in this report have internal consistency information on their scales. As explained in other parts of this report, scales are composed of items meant to measure a particular concept. Measuring internal consistency is often the first step in evaluating whether the items form a meaningful domain by determining whether they are cohesive (readers who would like a more extensive explanation with examples should refer to the Appendix).

However, Smith points out that internal consistency is only appropriate when the items are reflections of a non-tangible concept (called reflective for short). As an example, consider the concept "Supportive Environment." Although this might be an important concept to assess, one cannot measure it the same way one would measure temperature or weight. Instead, researchers must rely on a set of questions to approximate how supportive the environment is for children. One analogy for a reflective concept could be an art sculpture – to truly appreciate it, one must look at the sculpture from multiple angles.

³ The Weikart Center is a joint venture between the High/Scope Educational Research Foundation and the Forum for Youth Investment.

Similarly, to truly measure a reflective concept, one must examine it using a group of similar items (which provide the different angles). Theoretically speaking, researchers interested in developing a scale could probably generate hundreds of possible items to measure a particular concept. Of course, it is impractical to use them all, so researchers choose a manageable subset.

In contrast, internal consistency scores are not appropriate when concepts are formative, which means that a single concept is a composite of multiple, separate components (MacKenzie, Podsakoff, & Jarvis, 2005). A good analogy for this type of concept is a puzzle. To assess a formative concept, you need to gather all of the pieces and put them together. For example, imagine that we wish to measure “Program Resources.” Unlike a reflective concept, this type of concept is a composite of several important components (the puzzle pieces).

Our items would inquire about things like money, time, space, and number of staff members. Each of these resources may be an important component of the overall concept and are essential to include in the scale if we are to obtain a clear picture and complete the puzzle. Unlike reflective concepts, researchers cannot choose a subset of items from a large list of possibilities. Rather, each item is an important component to the whole. Because the concept is a composite of separate and potentially unrelated parts, the cohesiveness of the items is not important, and therefore internal consistency procedures are not appropriate (as stated in the Appendix, internal consistency measures the relatedness of items, which assumes that the items are reflective).

The Weikart Center has been reexamining the YPOA scales to assess whether they are reflective versus formative. Although this work is still in progress, the results will have important implications for how we think about evaluating both reliability and validity for observation-based metrics.

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.

	<i>Cost</i>	<i>Training Available</i>	<i>Estimated Time Necessary to Train Observers to Generate Reliable Scores</i>	<i>Estimated Minimum Observation Time Needed to Generate Sound Data</i>	<i>Data Collection, Management and Reporting Available</i>
<i>Assessing Afterschool Program Practices Tool (APT)</i>	Free*	Yes	4 hour training plus 2 program observations	1 afternoon (2-3 hours)	No
<i>Communities Organizing Resources to Advance Learning Observation Tool (CORAL)</i>	Free	No	2 days	3-4 hours	No
<i>Out-of-School Time Observation Tool (OST)</i>	Free	No †	8-18 hours, depending on experience	3 hours	No †
<i>Program Observation Tool (POT)</i>	\$300 Advancing Quality Kit	Yes	2.5-3 days	3-5 hours (for self-assessment)	No
<i>Program Quality Observation Scale (PQO)</i>	Free	No †	2 hours plus 2-4 observations & 2-4 time samples, depending on experience	1.5 hours observation & .5 hours time sampling	No †
<i>Program Quality Self-Assessment Tool (QSA)</i>	Free	Yes	2 hours ††	N/A	No
<i>Promising Practices Rating Scale (PPRS)</i>	Free	No †	2 hours plus 2-4 observations, depending on experience	2 hours	No †
<i>Quality Assurance System (QAS)</i>	\$75 Annual Site License	Yes	2-3 hours ††	1 afternoon (2-3 hours)	Yes
<i>School-Age Care Environment Rating Scale (SACERS)</i>	\$15.95 SACERS Booklet	Yes	4-5 days	3 hours	Yes
<i>Youth Program Quality Assessment (YPQA)</i>	\$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 NIOST 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%20Instrument.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.foundationinc.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 Quality⁴

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.

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

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

Individual Tool Descriptions

At-a-glance descriptions of the ten assessment tools are provided in the previous section. Here we offer more detailed descriptions. Each write-up 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

Purpose and History

The Assessing Afterschool Program Practices Tool (APT) is a set of observation and questionnaire tools designed to help practitioners examine and improve what they do in their after-school program to support young people's learning and development. It was specifically designed to examine those program practices that research suggests may be related to key youth outcomes (e.g., behavior, initiative, social relationships) and it is a core component of the Afterschool Program Assessment System (APAS).⁵

The research version of the APT (the APT-R) was developed in 2003-2004 for use in the Massachusetts Afterschool Research Study (MARS). Based on extensive field testing by grantees as well as some additional testing of the scales using MARS data, a more user-friendly self-assessment version of the tool was developed during 2005 for use by the Massachusetts Department of Elementary and Secondary Education 21st Century Community Learning Centers (21st CCLC) grantees and other programs interested in quality assessment. The self-assessment version is the focus of this description.

The instrument can be used to measure quality in a wide variety of program models that serve elementary and middle school students during the non-school hours. In addition to serving as a self-assessment tool, the APT defines desirable program practices in concrete terms that can be used to communicate with staff and others, help stimulate reflection and discussion regarding program strengths and weaknesses, guide the creation of professional development priorities and improvement goals and help gauge progress toward those goals.

The APT focuses on the experiences of youth in programs and is not intended to evaluate individual staff performance or produce definitive global quality scores

⁵ The APT was designed to address program practices that research suggests lead to youth outcomes measured by the Survey of Afterschool Youth Outcomes (SAYO) – an evaluation system developed by NIOST under contract with the MA Department of Education. The SAYO includes pre- and post-participation surveys for teachers and program staff and measures things like behavior in the classroom and program, initiative, engagement in learning, relations with peers and adults, homework, analysis and problem-solving and academic performance. For more information, see www.niost.org/training/APASbrochureforweb.pdf

for programs. While the APT includes a four-point rating scale, assigning ratings is not required; programs are encouraged to use the tool in ways that will yield the most useful information to guide program improvement.

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
- Youth participation

While it does address some broader organizational policy issues (e.g., connections with schools, staff-youth ratios) it was designed to focus primarily on things that program staff have control over and that are relevant across a range of different organizational contexts and facilities (e.g., schools, community centers).

The APT emphasizes some aspects of settings more than others and in particular places a strong emphasis on relationships, as research has shown that relationships have the greatest impact on youth outcomes. The primary focus is therefore on social processes – or interactions between people within the program. Several items help users observe and assess youth-adult relationships and interactions, as well as peer interactions and connections with families and school personnel. To a lesser extent than social processes, the APT also addresses programs' human and material resources and how those resources are organized or arranged within the setting.

In developing the APT, the authors reviewed relevant literature to identify program features that relate to outcomes and also looked at existing definitions and measures of program quality. One such definition commonly referenced in the field is the National Research Council's features of positive developmental

settings, a framework which itself is focused primarily on social processes. Items on the APT address each of the eight features identified by the National Research Council (2002).

Structure and Methodology

The 15 program features addressed by the APT are measured by one or both of 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.

Although the 15 program features drive the content of the tool, the APT-O is organized by daily routine. Five sections are intended to follow what the developers consider a typical program day. While these sections most closely reflect the daily routine in elementary and middle school 21st CCLC programs, the tool is designed to be flexible and users are encouraged to use whichever sections are most relevant in whatever order makes sense. These five sections include both informal program times (arrival, transitions, pick-up) and formal program times (homework, activities).

Within each section, sub-sections focus on particular practices and behaviors during those time periods (for example, sub-sections under “homework” include homework organization, youth participation in homework time, staff effectively manage homework time and staff provide individualized homework support). Finally, each sub-section includes between two and eight specific items that can be observed and rated.

An important structural aspect of the APT is its explicit connection to a specific youth outcome measurement tool – the Survey of Afterschool Youth Outcomes (SAYO). Programs can use APT findings to look at how they may be contributing to specific outcome areas included in the SAYO and users are provided with charts connecting particular APT items to specific outcome areas. Despite this linkage, the APT is also useful as a stand-alone tool.

<i>Program Feature</i>	<i>APT-O</i>	<i>APT-Q</i>
<i>Welcoming & Inclusive Environment</i>	✓	✓
<i>Positive Behavior Guidance</i>	✓	
<i>High Program & Activity Organization</i>	✓	
<i>Supportive Staff-Youth Relationships</i>	✓	✓
<i>Positive Peer Relations</i>	✓	
<i>Staff/Program Supports Individualized Needs & Interests</i>	✓	✓
<i>Staff/Programming Stimulates Engagement & Thinking</i>	✓	✓
<i>Targeted SAYO Skill Building/ Activities</i>	✓	✓
<i>Youth are Positively Engaged in Program/Skill Building</i>	✓	
<i>Varied/Flexible Approaches to Programming</i>	✓	✓
<i>Space is Conducive to Learning</i>	✓	
<i>Connections with Families</i>	✓	✓
<i>Opportunities for Responsibility, Autonomy & Leadership</i>		✓
<i>Connections with Schools</i>		✓
<i>Program Supports Staff</i>		✓

Recently, a new set of resources, the APT-SAYO Links, were developed as quick guides for practitioners to understand the research base connecting APT program practices and specific SAYO outcome areas.

The APT-O

The APT-O rating scale, should users decide to assign ratings to their observations, is a four-point scale designed to answer the question, “How true is it that this statement describes what I observed?” Definitions of each point on the scale differ slightly depending on whether you are observing a program or staff practice vs. youth behaviors. A detailed description of the rating scale as well as other rating options and considerations are included in the instruction manual. Some “conditional” items are included, which are only to be rated should they occur (e.g., when youth behavior is inappropriate, staff use simple reminders to redirect behavior).

Raters are asked to assign a 1-4 (or N/A) rating to each of the individual items within each sub-section. For most items, a specific description of what a “1” looks like is provided. The wording of the item itself constitutes a “4” since the question driving the ratings is, “How

true is this?” The instruction manual provides general guidance (not item-specific) for how to think about 2s (e.g., desired practice was only partially met, some minor evidence of negative expressions of the practice, or the practice is observed infrequently) and 3s (observed practice mostly reflected desired practice, or the desired practice was observed but perhaps not at all expected times). This year, NIOST will begin developing more specific anchors for 2s and 3s on the APT rating scale for each item. These new anchors will be field-tested, but not psychometrically tested this academic year.

The APT-Q

The APT Program Questionnaire (APT-Q) helps programs reflect upon the aspects of quality that are not necessarily observed, such as program planning, frequency of offerings, and connections with parents and schools. As is the case with the APT-O, flexibility is built into the questionnaire component of the tool. Users

<i>Arrival Time</i>	<i>How true?</i>	<i>Notes</i>
1. There is an established arrival routine that seems familiar to staff and youth.	1 2 3 4 N/A	
2. Activities are available to youth to become engaged in as they arrive (may include snack). (e.g., Wide variety of activities are available to arriving youth.) 1 = There are no activities available for arriving youth. Youth have nothing to do (e.g. Stand around waiting for staff to begin programming).	1 2 3 4 N/A	
3. Staff acknowledge children/youth when they arrive. (e.g., Offer a greeting, slap hands, ask “How’s it going?”, exchange hellos, etc.) 1 = Staff do not acknowledge any arriving youth.	1 2 3 4 N/A	
4. Staff engage in 1:1 conversations with youth. (e.g., Talk about youth’s day, ask about something they brought or made). 1 = Staff are not seen conversing or interacting with individual youth.	1 2 3 4 N/A	

APT Program Questionnaire Sections

1. How you plan and design program offerings
2. Your program offerings
3. How your program promotes responsibility, autonomy and leadership
4. How your program creates a welcoming and inclusive environment
5. How your program supports youth as individuals
6. How your program connects with families
7. How your program partners with schools to support youth
8. How your program supports and utilizes staff to promote quality

are encouraged to assign ratings only to the extent it is useful to do so and to review the various sections of the questionnaire before use to select those that best match a program's priorities. The APT-Q, which is divided into eight sections (see box), provides opportunities to rate the consistency and/or frequency of certain practices. It also provides lists of specific program practices that support various quality features (e.g., ways to create a welcoming and inclusive environment), encouraging users to check those that are in use in the program but at the same time offering a broad range of concrete, positive practices that can encourage program development and innovation.

The APT-Q includes three different rating scales. A four-point "how consistently" scale (rarely/never; once in a while/sometimes; often/a lot of the time; almost always/always) is used with the section focused on program planning and the use of specific planning practices. A six-point "how frequently" scale (rarely/never; a few times per year; about once per month; about once per

week; more than once per week; usually every day) is used for two sections that look at program offerings and to what extent the program promotes responsibility, autonomy and leadership. A simpler four-point "how frequently" scale (about once per year; several times per year; about once per month; weekly or more often) is used for the sections that address how the program connects with families and schools.

Technical Properties

The psychometric information that is available on the APT comes from the version used in the Massachusetts Afterschool Research Study (MARS), conducted by the Intercultural Center for Research in Education and the National Institute on Out-of-School Time (2005).⁶ The extent to which trained raters agree when observing the same program at the same time, or interrater reliability, was moderate and preliminary evidence for concurrent and predictive validity suggest the APT-R yields accurate information about the concepts it measures. As mentioned in the previous section, the anchors for the 2 and 3 ratings will be developed with the intent of improving interrater reliability. If funding for further testing comes through, NIOST will be re-testing interrater reliability.

While the current self-assessment version of the tool has no psychometric data, NIOST is currently seeking funding to conduct further psychometric testing of the APT and SAYO tools, including the extent to which the two tools can work together as an integrated measurement system and allow for practitioners to target key practices and track expected outcomes.

Interrater Reliability

Researchers examined interrater reliability for 78 programs in the MARS study and found that paired raters agreed on their ratings (within one score point) 85 percent of the time. If the range of response options for the research and self-assessment versions is similar, then we can expect, simply by chance, agreement between raters to be at least 62.5 percent, yielding a maximum kappa score of 0.60 (a high kappa is generally

⁶ The developers have conducted a detailed comparison of the two versions. Roughly half of the APT-R items in the current self-assessment version appear exactly as they were worded in the research version. Roughly one quarter of the original items were taken out, roughly one quarter were revised slightly.

considered 0.70). Although interrater reliability has not yet been established for the self-assessment version of the APT, existing data suggest that agreement was moderately better than chance.

Face Validity

The developers received extensive and systematic feedback on the APT, about usability as well as perceived validity. Twenty-six grantees representing over 100 program sites responded to a set of questions about the tool, most staff participated in focus groups and in-depth interviews were conducted with 12 grantees. Positive feedback from this range of key stakeholders suggested that the items make intuitive sense for measuring program quality. However, this is the weakest form of validity as it is not based on empirical evidence.

Concurrent Validity

MARS authors compared findings from the APT-R to observations of program characteristics and found that certain items were related to program and staff characteristics in expected ways. For example, programs that scored high on items related to staff engagement and engaging activities tended to have smaller group sizes, stronger connections with schools, a higher staff-to-child ratio and a higher percentage of staff with college degrees. Programs that scored high on items relating to youth engagement tended to be well-paced, organized with clear routines, had a higher staff-to-child ratio and a higher percentage of staff with college degrees. Better family relations were related to stronger connections between programs and parents and the community. Programs that offered high quality homework time tended to offer more project-based learning activities, were more organized with clear routines, had a higher staff-to-child ratio and had more staff that were certified teachers. NIOST's proposed study includes concurrent validity testing of the APT using the YPOA.

This evidence of concurrent validity should be regarded as preliminary because many items were not related to program characteristics. For example, youth engagement in programs was unrelated to smaller group sizes and engaging and challenging activities were not related

to programs being well-paced and organized with clear routines. Because the authors did not explicitly identify which relationships were most important and which findings ran contrary to their expectations, we cannot be certain that the observed findings indicate unequivocally strong concurrent validity.

MARS authors also examined the association between APT-R scales and five student characteristics that would be expected from theory and prior research: Students' improvement in their (1) behavior, (2) initiative, (3) homework, (4) relations with peers, and (5) relations with adults. Authors found that the APT-R Youth Engagement scale was related to all student characteristics in expected ways except Relations with Adults. Higher scores on APT-R Challenging Activities was related to lower scores on Relations with Peers, and higher scores on APT-R Relations with Families was marginally related to higher scores on Relations with Adults, but both APT-R scales were unrelated to other student characteristics. The APT-R Homework Time scale was unrelated to all student characteristics. MARS authors point out that the programs in their sample did not score high on Challenging Activities or Homework Time, indicating that the results could be different in a sample with more diverse programs on these dimensions. In addition, correlational evidence did not show any significant relationships between student characteristics and APT-R scales concerning Appropriate Space and Staff Engagement. Given these findings as well as keeping in mind that some, but not all, items used in the APT-R appear in the self-assessment version of the tool, this evidence should be regarded as preliminary, with additional evidence needed to firmly establish the concurrent validity of the APT.

Validity of Scale Structure

The authors created several scales in the APT-R using a statistical technique known as factor analysis. However the extent to which these findings can be generalized to the self-assessment version is unclear given the differences between the two instruments. Testing on the validity of the scale structure has been planned in NIOST's upcoming study.

User Considerations

Ease of Use

To date, the APT has been used by 36 school districts offering 21st CCLC programs in over 150 sites, and in a four-city pilot (Charlotte, Boston, Middlesex County, New Jersey and Atlanta). In response to its broad use, NIOST has developed many products related to the APT that make it both user-friendly and adaptable to a variety of systems and settings.

One such product is a package of tools, including the APT, the SAYO and a data management tracking system which tracks individual youth's program participation (an example of such a system is Kidtrax). When combined, these tools are referred to as the "Afterschool Program Assessment System (APAS)." The system can be used to support on-line data collection, management, analysis and reporting, allowing programs to link quality data from the APT and youth outcomes data from the SAYO with information on daily attendance and program participation in a comprehensive, flexible and integrated fashion.

NIOST has developed other products to support the use of the APT in conjunction with its other tools as part of a full assessment system (although the APT may be used as a stand-alone tool). For example, in addition to the SAYO, NIOST has created a series of SAYO-APT Links or "cheat sheets" to support programs' efforts to link quality and outcomes. The SAYO-APT Links describe each SAYO youth outcome area and correspond it with the related program practice found in the APT. Another product is a detailed, practitioner-friendly training notebook that guides programs in the use of the APT and related tools. Interested programs should contact NIOST for training options and costs; trainings are customized to meet the needs and interests of organizations.

In another development, NIOST is finalizing an online Youth Survey (SAYO-Y) designed to complement the APT and SAYO. The survey measures youth's program experiences in five areas (e.g., inclusive environment, choice & autonomy); youth's sense of competence in six areas (e.g., math, reading); and youth's future planning and expectations. This survey has been tested with over

6,000 Massachusetts youth participating in 21st CCLC programs and will be fully piloted again this year.

Flexibility is a hall-mark of the APT, so although the developers provide some guidance as to when to conduct observations, for how long, etc., they emphasize that the APT-O can be used in many different ways and that decisions about how many observers, how many observations and whether to use numerical ratings should be driven by what users intend to do with the data in the end. The general intention behind the design is for an observer to observe one full program session (typically a full afternoon), taking notes during the observation and using time immediately afterwards to complete all relevant sections including an open-ended "impressions" section at the end of the tool. Many programs using the APT as a self-assessment, however, have preferred to obtain at least two days of observation data. In its training sessions, NIOST guides participants in understanding how they may want to use tools, and, therefore, how they will tailor them to support their evaluation and assessment objectives.

Available Supports

Training on both the APT itself and processes for guiding program improvement is available through NIOST. Most recently, a two-day APAS training module has been developed to prepare site directors and others to use the SAYO and APT. Training for the APT, alone, is one full day. NIOST also provides a Quality Advisor training to help coaches and other technical specialists use the APT to work with programs. An online tutorials is also available to prepare sites to use the SAYO outcome tool.

In mid-2007, packaging and pricing information about training on the instrument became available for organizations that are interested but not already affiliated with the APT through statewide efforts in Massachusetts.

In The Field

The City of Cambridge, MA Agenda for Children is a city-wide out-of-school time initiative bringing together city departments, community-based organizations,

businesses, funders and residents to positively impact the lives of Cambridge youth. Specifically, the initiative works to improve access to and the quality of Cambridge out-of-school time programs. In pursuit of that goal, Cambridge Agenda for Children has been refining a multi-site program improvement process over three years, using the APT program improvement tool in that effort. Their Self-Assessment Support initiative supports programs to engage in observation and self-reflection, and take targeted action based on what they see.

The APT gives program coordinators, site coordinators and front-line staff a common language for talking about their goals and detailed descriptions of effective practice.

For More Information

Information about the APT is available online at:
www.niost.org/content/view/1572/282/
or www.doe.mass.edu/21cclc/ta

Contact

Kathy Schleyer, Training Operations Manager
National Institute of Out-of-School Time
Wellesley Centers for Women
Wellesley College
106 Central Street
Wellesley, MA 02481
781.283.2546
kschleye@wellesley.edu

Communities Organizing Resources to Advance Learning Observation Tool

Developed by Public/Private Ventures

Purpose and History

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.

Under the CORAL initiative, after-school programming was provided to elementary school aged children in five cities in California. In each of these cities, programming was different and consisted of a variety of activities ranging in focus from science-based programs to art and cultural enrichment programming. All CORAL programs included the common core element of Balanced Literacy programming.

The observation tool was used in two ways: first, to observe Balanced Literacy instruction in CORAL after-school programs, and second, to observe the integration of literacy programming in a variety of other activities including homework help and academic enrichment programs ranging in focus from science to art and cultural enrichment. 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.

The CORAL observation tool has five components:

- an *activity description form*, used to gather descriptive information on the observed activity prior to the observation;
- an *activity characteristic form*, completed during the observation, used to collect general information about the activity and type of instruction (i.e., length of time, number of participants, number of staff, teaching methods, etc.);
- an *activity checkbox form* that is divided into the five overarching categories for observation – this

is the primary method for recording information during the activity;

- the *activity scales form*, completed after the observation, that is used to rate each of the constructs observed in the activity;
- and the overall assessment component (completed after three observations for literacy programs and after two observations for non-literacy programs) which measures both aspects of the activity and participant improvement of skill areas.

Because the CORAL initiative emphasized best practices in youth development (including positive adult/youth relationships and ongoing youth participation), P/PV developed the CORAL tool based on an observation tool used for evaluation of the San Francisco Beacon initiative. P/PV also designed a similar observational tool containing the updated academic components from the CORAL tool for the Philadelphia Beacon initiative.

Content

The CORAL observation tool was designed to help researchers collect data through an ongoing program observation process to measure the connection between the quality of the program, fidelity to the Balanced Literacy Model, and the academic outcomes of participants. As a result, the main sections of the tool focus on five components of quality: adult-youth relations, effective instruction, peer cooperation, behavior management, and literacy instruction. Each aspect of quality has several elements of quality that are rated and captured in subcategories, which are called constructs. A number of these core constructs (such as behavior management, youth/adult relationships, peer cooperation, etc.) are relevant for both formal program settings as well as informal, adult supervised settings. The characteristics rated within each of the constructs are listed on page 38.

Because the focus of the CORAL observation tool is on fidelity to the Balanced Literacy Model – a structured literacy approach that uses a variety of modalities aimed at developing competent readers – it tends to focus on social processes and skill development in support of literacy gains, and less on program resources or the organization

<i>Adult-Youth Relationships</i>	<i>Behavior Management</i>	<i>Instruction</i>	<i>Literacy</i>	<i>Peer Cooperation</i>
<ul style="list-style-type: none"> • Adult support for the activity • General adult responsiveness • Emotional quality of the relationship 	<ul style="list-style-type: none"> • Appropriateness of behavioral demands • Adult management • Staff's inclusiveness of youth 	<ul style="list-style-type: none"> • Clarity • Organization • Motivation • Challenge • Connection to other material • Connection between youth & material • Cultural awareness • Responsiveness to English language learners 	<ul style="list-style-type: none"> • Literacy rich environment • Read aloud • Book talk/discussion/shout out • Writing • Independent reading • Skill development activities/games • Build vocabulary/spelling • Connections between youth & text 	<ul style="list-style-type: none"> • Cooperative activity

of those resources within a program. Literacy focused activities are assessed for their fidelity to the model and their association with change in participant interest and motivation. Non-literacy focused activities are assessed for their integration of literacy skills into the curriculum.

Structure and Methodology

The first three components of the CORAL observation tool – 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 *activity description form* is completed before (or after) the activity based on the information gleaned from the instructor in a 10-15 minute conversation. The *activity characteristics form*, where general information is recorded, is completed during the first ten minutes of the activity. The *activity checkbox form* is a running list of the observations of behavior and activity characteristics, and is filled out by the observer while the activity is ongoing. In the *activity checkbox form*, observers can choose between several examples of positive and negative behavior for each of the constructs above. If they observe a behavior not captured in the examples, it can be recorded in the notes section and

considered in the scoring after the observation is completed. Observations are conducted over a 90-minute period (in CORAL, the literacy activity took place over a 90-minute period, so the observation took place over its entire duration in order to be able to assess fidelity).

Each construct is based on a five-point rating scale, with 1 representing the lowest score (definitely needs improvement) and 5 the highest (outstanding). The CORAL developers suggest that 5s be awarded sparingly as the intent is to indicate that there is no room for improvement. Meanwhile, the 1 rating can be given for several reasons, including observed negative staff behaviors or due to an activity not fitting the appropriate construct. Activity ratings are assigned after each 90 minute observation for cumulative observations of three for a literacy program and two for a non-literacy activity.

After the 90 minute observation, observers complete the activity scales form within 24 hours of each observation, including the descriptive narrative. The *checkbox form* does not translate into a one-to-one score on the scales form. Instead observers are required to consider the activities recorded on the *checkbox form* along with the

Q8. Challenge (++)	Q8. Challenge (--)
<p>The staff:</p> <ul style="list-style-type: none"> encourage youth to push beyond their present level of competency. try to sustain the motivation of youth who are discouraged or reluctant to try. continuously move to the next step as soon as youth progress. reinforce and encourage youth's efforts in order to maintain their involvement. 	<p>The staff:</p> <ul style="list-style-type: none"> discourage youth who tried to push beyond their present level of competency. miss opportunities to sustain the motivation of youth who are discouraged or reluctant to try. miss opportunities to move to the next step as soon as youth progressed (e.g., pace was too slow) do not reinforce or encourage youth's efforts in order to maintain their involvement.
N/A because:	

duration/frequency of the observed behavior, quality of the behavior and importance of the behavior to the activity when deriving a score for each construct. Additionally, observers complete descriptive narrative summaries (using sample narratives as a guide) which contain the most informative aspects of the activity.

Upon completion of the required series of observations, the overall assessment form is used to rate the overall quality of the activity. The form contains 11 narrative questions in which observers describe the strengths, weaknesses, areas of improvement for the activity, cultural awareness of the instructor, modifications for varying linguistic needs, and the classroom environment. The observers are also asked to record the improvements they observed. The question contains a pre-selected list of skills which range from academic to visual arts and performance.

The developers suggest three or four hours for completing the rating scales, related narratives and the overall assessment.

For their research purposes, Public/Private Ventures additionally required observers to write a narrative description for each component and used the descriptions as part of their quantitative analysis and as a method of compiling "best practices" that exemplified specific constructs. In addition, observers were asked to identify in a narrative which aspect of each of these skills participants showed improvement.

Technical Properties

Technical evidence for the CORAL observation tool comes from its use during the evaluation of the CORAL Initiative, a two-year study that included 56 observations in 23 after-school programs in the first year and 43 observation in 21 after-school programs in the second year. The 90-minute literacy activities were observed two to four times each during the first year of the evaluation, and a minimum of three times each during the second. Evidence is drawn from study's initial report after the first year (Arbreton, Goldsmith & Sheldon, 2005) as well as the final report, which draws on evidence from both years (Arbreton et al., 2008). Data from the non-literacy activity observations were used to create qualitative profiles of the activities, and no statistical analysis was conducted. For the literacy activities in the CORAL evaluation, statistical analysis was conducted to identify the relationship between program quality and participant academic gains. The technical properties described below pertain to P/PV's analysis of literacy observation data only.

Some users may be interested in summarizing data from the CORAL into scales. In their analysis from the CORAL initiative (Arbreton, Goldsmith, & Sheldon, 2005; Arbreton et al., 2008), the developers created four scales using items from the instruments' Activity Scales Form: (1) Adult-Youth Relation (average of items 1 through 3); (2) Instructional Quality (average of items 5 through 8); (3) Group Management (average of items 16 and 17);

(4) Connection Between Youth and Activities (average of items 9 and 10).

Developers also created an Overall Lesson Rating, which is an average of scores from three scales (Adult-Youth Relation, Instructional Quality, & Group Management), and several items: Read Aloud (item 19), Book Talk (item 20), Writing (item 21), Independent Reading (item 22), & Connection to Youth (item 10).

Score Distributions

Items and scales should be able to detect meaningful differences across settings, and therefore should exhibit a range of scores. Score distributions were examined for the four domain-specific scales as well as six individual items representing the Balanced Literacy strategies measured in the CORAL tool. Most items exhibited good score distributions, although the item “Skill Development Activities” (item 24) was on the low end of the scale for both years of measurement (average scores were 1.5 and 1.8, respectively, on a scale of 1 to 5).

Internal Consistency

Responses to items comprising scales should be highly related, suggesting that the items form meaningful domains. The internal consistency of the Overall Lesson Rating was quite strong with an alpha of .94. Of the four scales measuring specific domains from the CORAL Initiative, three (Adult Support, Instructional Quality, and Group Management) exhibited excellent internal consistency with Cronbach’s alpha ranging from .84 to .88 (exceeding the recommended value of .70), suggesting that the scales are cohesive and composed of related items. However, the scale Connection to Youth and Activities was less cohesive (alpha=.54), suggesting that the two items composing this scale are only moderately related and more items may be needed to fully capture this domain.

Validity of Scale Structure

An analysis that examines scale structure validity for a single scale tests whether the scale truly measures a distinct and coherent domain (rather than several different concepts or domains). Developers examined whether several items and scales from the CORAL

tool could be combined to represent an Overall Lesson Rating. Findings indicated that the Overall Lesson Rating represents a cohesive summary of multiple domains from the CORAL measure. The developers did not examine scale structure validity for other scales measuring specific domains (such as Instructional Quality).

Predictive Validity

If the CORAL tool truly measures effective literacy strategies, we can expect that scores from the literacy strategy items will be related to gains in reading and English language skills. To examine the instrument’s predictive validity, authors examined the relationship between scores on the CORAL tool with outcomes of 234-383 children across the two years of the study.

As discussed in the CORAL initiative reports (Arbreton, Goldsmith, & Sheldon, 2005; Arbreton et al., 2008), authors classified programs into five literacy profiles based on how the programs scored on the six items measuring Balanced Literacy strategies. Readers who are interested in more information on how the literacy profiles were created should consult the CORAL initiative reports cited in this document. At the end of the first year, children who attended programs with better literacy profiles had greater reading improvements on the Informal Reading Inventory (IRI) but were not more likely to have positive outcomes on the California Standards Test- English-Language Arts (CST-ELA). One reason why literacy profiles may have predicted better scores on the IRI but not the CST-ELA could be because the IRI focuses largely on reading abilities and comprehension, whereas the CST-ELA is somewhat broader and also incorporates writing skills and word analysis. Classroom practices (as measured by the four scales) did not by themselves predict reading improvements in either the IRI or the CST-ELA, over the first year of the evaluation.

At the end of the second year, the authors created a scale called the Overall Lesson Rating that is a combination of items measuring literacy strategies as well as classroom practices. The Overall Lesson Rating predicted more positive outcomes on the CST-ELA (scores on the IRI were not examined in the second

year). The findings from both years suggest that the CORAL tool successfully measures quality in programs with a strong literacy component.

User Considerations

Ease of Use

Although the CORAL tool can be used by anyone, it was designed exclusively for research purposes and has not been specifically adapted for practitioner use at this time. However, the tool does contain detailed instructions for conducting observations and completing the forms.

Available Supports

At this time, Public/Private Ventures is not offering training on use of the CORAL observation tool, though a two-day training was offered to the observers participating in the evaluation study. This training included a review of the observation materials, mock observations and write-ups on day one, and a field observation in which a new observer was paired with an experienced guide on day two. The trainee and guide compared notes to develop consistency. Trainees received ongoing monitoring and support.

In the Field

In the 2004-2005 school year, the CORAL observation tool was used as part of an evaluation of the programs' implementation of the Balanced Literacy Model. The evaluation was conducted by Public/Private Ventures. Live observations of the participants' experiences, the impact of balanced literacy on the academic gains of participants, and fidelity to the Balanced Literacy Model were at the center of the research. Observations were conducted during the program participants' third-grade and fourth-grade years, and each participant was observed three times by an observer assigned from a pool of observers.

In analyzing the data, researchers found that although the link between quality and academic gains was inconclusive, the findings indicate that the participants with the greatest academic gains were those that participated in higher quality programs. Additionally, all

youth, including those reading below grade level, had greater gains when in the higher quality programs than did their counterparts in lower quality programs. In the second year of observations, the researchers observed that there was consistency in the implementation of the literacy model as well as higher quality implementation. In the same time period, researchers also observed reading gains that were 39 percent higher than the gains recorded in the first year.

For More Information:

Information about the CORAL observation tool is available at:

www.ppv.org/ppv/initiative.asp?section__id=0&initiative__id=29

Contact:

Amy Arbreton, Senior Research Fellow
Public/Private Ventures
Lake Merritt Plaza
1999 Harrison Street, Suite 1550
Oakland, CA 94612
510.273.4600

Out-of-School Time Program Observation Tool

Developed by Policy Studies Associates, Inc.

Purpose and History

Policy Studies Associates (PSA) developed the Out-of-School Time Program Observation Tool (OST Observation Tool) over a five-year period, in conjunction with several research projects related to after-school programming, including a major study of promising practices in high-performing programs (Birmingham, Pechman, Russell & Mielke, 2005). A third edition of the tool, revised in 2008, was reviewed for this compendium. The instrument was recently used in studies of the New York City Department of Youth and Community Development's Out-of-School Time (OST) Programs for Youth and of the New Jersey After 3 Initiative.

The tool was developed with research goals in mind – in particular the desire to collect consistent and objective data about the quality of after-school activities through observation. Its design is based on a couple of 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 development outcomes.

The OST Observation Tool can be used in varied after-school contexts including school- or center-based programs and with youth participants that are in kindergarten through 12th grade. While the tool can provide program staff with a framework for observing and reflecting on their practice, it was developed for, and thus far has primarily been used for, research purposes. In its current design, it is not intended to be used to assign overall quality scores for programs or staff.

Content

The OST Observation Tool was designed to provide researchers and other users with a framework for observing essential indicators of positive youth development. It focuses on three major components of programs: activity type, activity structures and interactions between youth and adults and among

youth. The first section captures a range of in-depth information about the type of activity being observed and the skills emphasized through that activity; the remainder focuses on what the youth development literature points to as critical components of programs.

Because of its developmental grounding and its focus on what young people experience inside of programs, the OST Observation Tool has an activity and program-level focus and does not address organizational issues related to management, leadership or policy. The primary focus is on social processes – including relational issues and many items that speak specifically to instruction and learning. Beyond one item related to materials, the instrument does not focus on program resources or the organization of those resources within the setting.

The content of the OST Observation Instrument aligns very closely with the SAFE framework (Durlak and Weissberg, 2007) which outlines features of programs that contribute to positive outcomes for youth in out-of-school time programs. SAFE refers to out-of-school time activities that are:

- **Sequenced:** Content and instruction are designed to increasingly advance skills and knowledge and help youth achieve goals;
- **Active:** Activities lend themselves to active engagement in learning;
- **Personally Focused:** Activities strengthen relationships among youth and between staff and youth;
- **Explicit:** Activities explicitly target specific learning or developmental goals.

The 2008 version of the tool was updated to fully align with the SAFE framework. The changes are most obvious in the reorganization of the qualitative portion of the instrument in which observational notes are recorded and synthesized. This framework was also used in the evaluators' analyses of their Year 2 findings for the New York and New Jersey studies which demonstrated that the OST indicators map well to the SAFE framework.

Additional changes between the second and third editions involve inclusion of the academic and technology features of programs. This new section features items related to literacy, math instruction and the use of technology, guiding users to note the presence or absence of activities that meet literacy, math or technology goals.

Structure and Methodology

The first part of the instrument, which focuses on activity type, provides observers with detailed definitions for documenting:

- Type of activity (e.g., tutoring, visual arts, music, sports, community service)
- Type of space (e.g., classroom, gym, library, auditorium, hallway, playground)
- Primary skill targeted (e.g., artistic, physical, literacy, numeracy, interpersonal)
- Number and education level of staff involved in the activity
- Environmental context (e.g., supervision, space, materials)
- Number, gender and grade level of participants

In addition, the instrument provides detailed descriptions of each of the four SAFE features for rating purposes. The above observations are recorded on a coversheet that also includes other basic information about the observer, program, date, time, etc.

The remainder of the tool addresses five key youth development “domains” including relationships (youth- and staff-directed are considered separately), youth participation, skill building and mastery, and activity

content and structure. Each domain is subdivided into four to seven specific indicators or practices. For each indicator, a detailed “exemplar” is offered to guide ratings. For example:

- **Domain:** Relationship-building: Youth.
- **Indicator:** All or most youth are friendly and relaxed with one another.
- **Exemplar:** Youth socialize informally. They are relaxed in their interactions with each other. They appear to enjoy one another’s company.

The rating scale in the OST Observation Instrument asks users to assess the extent to which each indicator is or is not present during an observation. While the developers have experimented with both three- and five-point scales in various studies, the third edition of the instrument uses a seven-point rating scale which gives more room for capturing subtleties, where 1=not evident and 7=highly evident and consistent (see below). A “5” rating is considered basic quality. Observers are instructed to first select the odd number that most closely reflects the level of evidence observed and then, if necessary, to move up or down to the adjacent even number if that more accurately reflects the presence of the indicator within the activity.

Developers of the OST Observation Tool have structured it flexibly so that users can arrange scales differently for different purposes. Although definitive rules for constructing scales do not exist, in their report on the validation study (Pechman et. al., 2008), the authors present four different methods for creating scales. The four sets of scales were used in different studies across several years and were each guided by separate theories and evaluation questions. The authors present the scale sets as different options for users to summarize data

— 1 —	— 2 —	— 3 —	— 4 —	— 5 —	— 6 —	— 7 —
Exemplar is not evident		Exemplar is rarely evident		Exemplar is moderately evident or implicit		Exemplar is highly evident and consistent

from the OST. Users who are interested in summarizing data should refer to the report of the validation study for information on which items compose specific scales.

The first scale set includes four scales: (1) Youth relationship-building and participation, (2) Staff-youth relationships, (3) Skill building and mastery and (4) Activity content and structure. These scales were used in a study of Shared Features of High-Performing After-School Programs conducted on behalf of The After-School Corporation and the Southwest Educational Development Laboratory (Birmingham et al, 2005)

The second scale set is similar and also includes four scales: (1) Youth relationship-building, (2) Staff-relationship building, (3) Instructional methods and (4) Activity content and structure. These scales were used in the first year of the evaluation of the New Jersey After 3 initiative (Kim et al., 2006).

The third scale set includes three scales: (1) Relationships, (2) Instructional strategies and (3) Activity content and structure. These scales were used in the first year of the evaluation of the New York City Department of Youth and Community Development's Out-of-School Time Programs for Youth initiative (Russell et al., 2006).

The fourth scale set is based on the SAFE framework which emphasizes activities that are **S**equenced, **A**ctive, **F**ocused and **E**xplicit, as described by Durlak and Weissberg (2007). The four scales in this set correspond to these four SAFE domains. These scales were used in the second year evaluations of both the New Jersey After 3 initiative and the New York City Department of Youth and Community Development's Out-of-School Time Programs for Youth initiative (Walking Eagle et al., 2008; Russell et al., 2008).

Technical Properties

Psychometric information presented here for the OST Observation Instrument comes from the three studies previously mentioned (Shared Features of High-Performing After-School Programs, New Jersey After

3, and the New York City Department of Youth and Community Development's Out-of-School Time Programs for Youth initiative).

Score Distributions

Pechman and colleagues (2008) examined the score distributions of all scales in each of the four scale sets. The authors generally found good variability in the scores across raters' observations of 159 to 238 activities in 10 to 15 programs observed across the three studies. One exception was the Active scale from the SAFE scale set. The average score for this scale was somewhat low (1.9 on a scale of 1 to 7), making it difficult to determine whether the scale has difficulty capturing differences across programs or most programs simply do not keep youth very active. However, findings suggest that all other scales capture meaningful differences across a variety of activities and programs.

Interrater Reliability

Observers using this instrument reached high levels of agreement. Examining data from five assessments across three separate studies, pairs of researchers co-observed between 19 to 40 activities within 10 to 15 programs at each assessment. Using Pearson and intraclass correlations, researchers examined the interrater reliability for the overall score as well as the average score within each of the instrument's five domains. When available, intraclass correlations were generally above the recommended value of .50, indicating strong agreement. Strong agreement was also supported by the Pearson correlations, which were close to or above the recommended value of .70.⁷ Therefore, these findings suggest that trained raters can achieve high overall agreement across each of the five domains.⁸

Internal Consistency

Internal consistency was strong for all scales across the four sets with Cronbach's alphas exceeding the recommended value of .70. These findings suggest that each of the scales' items are highly related and form meaningful domains. The first scale set summarized

⁷ The one exception was for the Activity Content and Structure domain, which was lower for one out of the five assessments, but agreement on the other four assessments was strong.

⁸ The interrater reliabilities for specific items and scales were not reported and could be lower.

TASC study data collected from 173 independent observations in 10 programs and had alpha levels ranging from .73 to .88. The second scale set summarized New Jersey After 3 data collected from 179 independent observations in 10 programs and had alpha levels ranging from .81 to .83. The third scale set summarized New York City OST study data collected from 238 independent observations from 15 programs and had alphas ranging from .80 to .87. The fourth scale set summarized both New York City OST study data and New Jersey After 3 data collected from a combined total of 358 observations and had alpha levels ranging from .84 to .88.

Concurrent Validity

OST developers examined the concurrent validity of the third scale set drawing on 1,444 youth surveys from the DYCD OST initiative in New York City and the New Jersey After 3 initiative. Specifically, using Spearman's Rho rank order correlation coefficients, researchers examined the associations between the OST Relationships, Instructional Strategies, and Activity Content and Structure scales with responses from a separate youth survey on Interactions with Staff, Interactions with Peers, Sense of Belonging, Exposure to New Experiences, and Academic Benefits.

Higher scores on the Relationships scale were related to higher scores for Exposure to New Experiences, Interactions with Peers, and Interactions with Staff for both years of the study, as well as higher scores for Sense of Belonging in the first year and Academic Benefits in the second year. Higher scores on the Instructional Strategies scale were related to higher scores on Exposure to New Experiences, Sense of Belonging, and Interactions with Staff in the first year of the study but not the second. Instructional Strategies was not related to Interactions with Peers or Academic Benefits, and the Activity Content and Structure Scale was not related to any youth experiences from the youth survey in either year.

Based on these findings, the available concurrent validity evidence is mixed. In addition, concurrent validity evidence does not currently exist for the other scale sets.

Validity of Scale Structure

An analysis examining scale structure validity tests whether items forming multiple scales truly measure distinct and coherent domain as expected. The instrument's developers conducted a factor analysis to examine the structural validity of the fourth scale set. Findings suggested that the items could be categorized into the four SAFE domains, although the Sequenced and Explicit domains were moderately related, suggesting that they are not entirely distinct from one another. In addition, the Active domain appeared to be a combination of two distinct categories, suggesting that this domain is not completely cohesive.

Although the instrument has some evidence of scale structure validity for the SAFE domains, evidence for the other three scale sets is currently unavailable.

User Considerations

Ease of Use

While the OST Observation Instrument is available online and is free for anyone to download and use, it is important to recognize that it was developed with primarily a research audience in mind. The introduction to the tool includes an overview and review of basic procedures for conducting observations and completing the form, but the materials have not been tailored for practitioners at this time and use language (e.g., sampling, reliability) that may not be accessible to some audiences.

Its developers consider the OST Observation Tool to be highly efficient to complete in the field. Users observe 15 minutes of an activity and score it immediately in less than five minutes. Users are advised to observe a total of 8-10 activities over at least two afternoons (or approximately three hours of program observation) to adequately sample program offerings. Additional guidance about how to organize observations on site, sample activities appropriately and manage multiple observers is provided in the instrument's procedures section.

Available Supports

At this time, training related to the OST Observation Instrument is limited to individuals involved in a

specific study that employs the instrument. Data collectors participate in trainings that provide a detailed overview of the instrument, its indicators and the theoretical framework. Following a review of the operational definitions for each category and group of indicators, researchers participate in practice rating sessions using video-taped samples of after-school activities to build interrater reliability prior to fieldwork. Additional reliability checks are conducted in the field and in follow-up meetings to ensure common interpretation of terms and items.

Researchers typically use the observation data collected with the OST instrument in conjunction with supplementary (but not formally linked) measures such as interviews, surveys and focus groups. As research continues, validity data will become available about the relationship between program quality features and youth outcomes as measured by some of these other instruments.

In The Field

In 2005, the New York City Department of Youth and Community Development (DYCD) contracted with PSA to conduct a comprehensive evaluation of its 536 OST programs serving 69,000 participants under its Out-of-School Time Programs for Youth initiative. Participating service providers, serving all grade levels, operated under one of three funding mechanisms: a) Option I, targeted toward a general pool of service providers operating programs in neighborhoods throughout New York City; b) Option II, for programs using a 30% private funding match; and c) Option III, for programs operated in collaboration with the Department of Parks and Recreation and offered at Parks sites.

The OST Observation Tool was used in a sample of 15 of the 536 OST programs as part of the overall evaluation. The evaluation combined this sampling data with other data sources, including a participation database and program director surveys to round out the picture. The first-year evaluation findings identified avenues for improving the effectiveness of OST programming in several areas. For example,

although programs successfully enrolled students in the first year, they struggled to maintain high youth participation rates, suggesting a need to establish program policies and activity offerings that encouraged regular participation. Additionally, while programs in the first year consistently provided safe and structured environments for participants, they experienced challenges in delivering innovative, content-based learning opportunities that engaged youth.

PSA's second year findings centered on evidence of programs' efforts to improve program quality and scale. The findings suggest that OST programs increased both their enrollment and participation rates. Programs scaled up enrollment from 51,000 youth in the previous year to serve more than 69,000 youth throughout New York City. Rates of individual youth participation also increased substantially compared to Year 1, indicating that programs were successfully recruiting and retaining participants. In addition, programs reported that they improved the quality and capacity of their program staff through improved hiring and professional development opportunities.

In year three, the evaluation will continue to collect data from OST programs to explore the associations among program-quality features, youth participation patterns, and youth outcomes.

For More Information

The OST Observation Instrument is available online at: www.policystudies.com/studies/youth/OST%20Instrument.html

Contact

Christina Russell or Ellen Pechman
Policy Studies Associates
1718 Connecticut Avenue, NW
Washington, DC 20009
202.939.9780
crussell@policystudies.com

Program Observation Tool

Developed by the National AfterSchool Association

Purpose and History

The Program Observation Tool is the centerpiece of the National AfterSchool Association's program improvement and accreditation process and is designed specifically to help programs assess progress against their Standards for Quality School-Age Care. The instrument was developed by the National AfterSchool Association (NAA) and the National Institute on Out-of-School Time in 1991 and was based on the Assessing School-Age Child Care Quality Program Observation Instrument developed by Susan O'Connor, Thelma Harms, Debby Cryer and Kathryn Wheeler. The instrument was revised in 1995 and piloted between 1995 and 1997. Additional revisions were then made before NAA's accreditation system became active in 1998.

The NAA Standards, which the Program Observation Tool is based on, are meant to provide a baseline of quality for after-school programs serving children and youth between ages 5 and 14. They are intended for use in group settings – primarily school and center-based – where children participate regularly and where the goal is supporting and enhancing overall development.

Rooted in the frameworks of the early childhood and school-age care fields of the early 1990's, the instrument and the NAA standards reflect much of the thinking of the time, particularly in terms of licensing and monitoring, and have been used as part of a seven-step accreditation process for the past decade. Pre-dating the creation of the federal 21st Century Community Learning Centers program, the NAA standards have played a significant role in the field and have been adopted by a range of programs and systems across the country. There are now 20,000 copies of the standards book in print and over 500 programs across the country are in some stage of the accreditation process.

In 2008, NAA shifted away from its role as an accrediting body and is now offering accreditation through the Council on Accreditation. NAA will complete the accreditation process through the end of 2009 for all agencies that applied and were at some stage of the process before September 2008. For agencies applying

after September 2008, a program relationship manager from the Council will assist them through the process.

The Program Observation Tool will still be available for agencies interested in using it for self-assessment and improvement purposes (as has always been the case for agencies not seeking accreditation). NAA will retain the rights to the standards and materials, and continue to provide supports for technical assistance.

Content

The Program Observation Tool measures 36 "keys of quality" that are organized into six categories. Five of those categories are considered observable and are assessed primarily through observation: human relationships; indoor environment; outdoor environment; activities; and safety, health and nutrition. The sixth category – administration – is assessed through questionnaire.

Because of NAA's commitment to supporting child development in a holistic way, the instrument measures a range of social processes – how children and staff within the setting interact. Because of the link to accreditation, it also focuses quite a bit on program resources and the arrangement (spatial, social and temporal) of those resources within the program. Unlike some of the other tools in this compendium, the Program Observation Tool also addresses program policies and procedures that are believed to influence quality.

The Program Observation Tool pre-dates the National Research Council's features of positive development settings framework (2002) by over a decade and draws more heavily on the early childhood literature than the youth development literature. However, it does address many of the NRC features, placing the least emphasis on "support for efficacy and mattering" and "skill building opportunities."

Structure and Methodology

The five quality categories that are the focus of the Program Observation Tool are measured using one overall 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 (included in the Standards, but not the Observation Tool) are assessed separately, through questionnaire/interview.

The rating scale used throughout the Program Observation Tool (see example below) is intended to capture whether each indicator is true all of the time (3), most of the time (2), sometimes (1) or not at all (0). Although specific descriptions of what a 0, 1, 2, or 3 looks like for each indicator are not provided, between one and eight descriptive bullet statements are included under each indicator to clarify meaning.

Space is provided for observers to take notes on each indicator. At the bottom of each page, observers are encouraged to total their numerical scores for each quality key to achieve an overall program rating. Tally sheets and instructions are provided for multiple observers to reconcile and combine their scores. In order to achieve accreditation, there are two “weighted” categories – program/activities and safety/nutrition in which programs must meet a certain threshold in order to be accredited.

6. Children and youth generally interact with one another in positive ways.		
Guiding Questions: Do children seem to enjoy spending time together? Do they talk about friends at the program? Do they tend to include others from different backgrounds or with different abilities in their play?		
	<i>Comments</i>	<i>Rating</i>
a. Children appear relaxed and involved with each other. <ul style="list-style-type: none"> Group sounds are pleasant most of the time. 		0 1 2 3
b. Children show respect for each other. <ul style="list-style-type: none"> Teasing, belittling or picking on particular children is uncommon. Children show sympathy for each other and help each other. 		0 1 2 3
c. Children usually cooperate and work well together. <ul style="list-style-type: none"> Children willingly share materials and space. They suggest activities, negotiate roles and jointly work out rules. Children include others with developmental, physical or language difference in their play. Children often help each other. There is a strong sense of community. 		0 1 2 3
d. When problems occur, children often try to discuss their differences and work out a solution. <ul style="list-style-type: none"> Children listen to each other’s point of view and try to compromise (e.g. if two children want to use the same equipment, they may decide to take turns as a solution). Children know how to solve problems. Their solutions are generally reasonable and fair. They do not try to solve disagreements by bullying or acting aggressively. 		0 1 2 3

Technical Properties

Although no psychometric evidence is available on the Program Observation Tool itself, there is information available about the ASQ (Assessing School-Age Childcare Quality), from which the POT was derived. Users should note that the ASQ's psychometric properties may not be completely consistent with those of the POT.⁹ Overall, evidence for interrater and test-retest reliability is strong for the ASQ, meaning the assessments of the same program practices by different observers are consistent and assessments are stable over time. Following revisions to the scales, evidence of internal consistency, or the degree to which items fit together in meaningful ways, was strong. Validity data are limited, although preliminary evidence for concurrent validity suggests the instrument may yield accurate information about the concepts it measures.¹⁰

The field study which provides psychometric support for the ASQ involved a sample of 40 after-school programs in Massachusetts and North Carolina (Knowlton & Cryer, 1994). Two versions of ASQ scales were examined: original and revised. The revised version's scales are comparable to those in the POT: Human Relationships, Indoor Environment, Outdoor Environment, Activities, and Safety, Health and Nutrition. Of the original scales, only two overlapped with the POT, namely human relationships and activities. When appropriate, we state which set of scales exhibits specific properties.

Interrater Reliability

To examine interrater reliability, paired raters evaluated 40 programs using the measure. ASQ indicators are organized into 21 items and those items are further organized into five scales. Knowlton and Cryer examined agreement among raters at both the item and scale levels. The kappa statistic measures the degree to which raters agree and corrects for cases where raters agree simply by chance. All items had kappa scores above .70, generally considered the threshold for high agreement. The authors also computed intraclass correlations and

all of the ASQ original scales and total score were near or above .70, showing good agreement on these scores.¹¹ However, because only the original ASQ scales, not the revised versions, were examined, we can generalize only for those scales that are similar (Human Relationships, Activities – and the total score).

Test-Retest Reliability

Ideally, instruments should be able to assess major changes over time but should exhibit stability in scores across multiple assessments in the short-term. For the ASQ, 25 programs were reassessed two weeks after their initial assessment to determine the instrument's test-retest reliability. Knowlton and Cryer (1994) found that all items demonstrated acceptable stability, with kappa scores above .70.¹² The authors also computed intraclass correlation coefficients to examine stability of the original scales and total score over time. All scales and total score were above 0.70, but we can only generalize to the scales that overlap with the POT – Human Relationships and Activities – and the total score.

Internal Consistency

To determine whether items within the scales fit together in meaningful ways, Knowlton and Cryer examined the internal consistency of the original scales and the total score by computing a statistic called Cronbach's alpha. The alpha for one of the original scales (Safety) was very low, so the authors revised the scales (and the revisions more closely match the POT). Results from the revised scales and the total score demonstrated good internal consistency, with alphas near or above the recommended cutoff of .70.

Convergent Validity

To determine the extent to which the ASQ yields accurate information about the aspects of programs

¹¹ Readers should note that Knowlton and Cryer also looked at the interrater reliability of the individual indicators that composed the items. Many indicators exhibited poor agreement. However, summing the indicators into items creates more reliable measures because it cancels out some of the measurement problems. For this reason, users should evaluate programs based on the items and scales, not the individual indicators.

¹² Similar to tests on interrater reliability, the authors found that 40 percent of the indicators had poor short-term stability. However, the measurement problems associated with individual indicators likely cancel out when creating an item score. Again, users should examine the items and scales, not the indicators, when evaluating programs.

⁹ There are slightly more indicators in the ASQ (84) than in the POT (80). It is unclear how many indicators are identical or similar.

¹⁰ The technical section only evaluates evidence from the observational portion of the instrument, not the administration questionnaire.

it is supposed to measure, Knowlton and Cryer (1994) compared the ASQ scores for 11 programs with subjective ratings by experts. Specifically, two experts ranked a set of programs in terms of overall quality within each of the five original ASQ domains using their own criteria. Using what is called the Spearman correlation, ASQ rankings were moderately to strongly related to the expert rankings, with the exception of the Safety and Health and Nutrition areas. This validity evidence should be regarded as preliminary, based on the small number of programs and experts included in the analysis and the fact that estimates were computed on the original, unrevised ASQ scales.

User Considerations

Ease of Use

The Program Observation Tool and NAA standards were developed with significant input from practitioners, resulting in accessible language and a user-friendly format.

Programs wishing to undertake the accreditation process can contact the Council on Accreditation (see contact information). For self-assessment purposes, observing the program and scoring the full instrument takes roughly 3-5 hours. The self-study manual provides very detailed guidance to program directors and staff on how long and how much of the program to observe, how to determine ratings and how to combine scores from different raters.

The observation tool is one of a package of products related to accreditation – the Advancing and Recognizing Quality Kit – which includes the standards book; the guide to program accreditation; self-study manuals that include the observation tool as well as staff, family and child/youth questionnaires; and a training video. The team leader’s manual walks program directors or staff through the various steps of the accreditation process in detail and includes specific tools for developing an action plan for improvement based on observational data. As a package, these resources cost approximately \$300. There are additional costs related to the full accreditation process.

Available Supports

It is important to reiterate that while this summary has focused specifically on the Program Observation Tool, that instrument is just one piece of an integrated set of resources related to self-study and accreditation. NAA offers training that covers the Program Observation Tool through its the day-long Endorser Training (NAA recommends two and a half days of training in order to ensure reliability). Some NAA state affiliates offer local training related to the instrument for programs interested in using it for self-assessment and improvement.

In the Field

The University of Missouri-Adventure Club is a district-wide after-school initiative for elementary school students in Missouri’s Columbia public school district. The National Afterschool Association’s standards and observation tool, as well as the larger Advancing School-Age Quality (ASQ) process within which these are embedded, serve as the organizing framework for Adventure Club’s 18 programs. Institutionalization of the standards has resulted in a common language and understanding of program quality that spans the individual staff, program and cross-site levels.

The Program Observation Tool is used by each of the 18 programs several times each year and is a core piece of the new staff orientation process, which includes conducting and discussing a program observation with more senior colleagues. Line-staff are well-versed in the 36 “keys of quality,” and each week during cross-site directors’ meetings one key is the focus of in-depth discussion.

In addition to regular observations – by staff, administrators and parents, each program has an ASQ team made up of these stakeholders (parents, staff, administrators and sometimes children). Teams meet monthly or bi-monthly to review new observation data and revisit the program’s improvement plan. “This is a continuous process – it doesn’t start and stop each year. Each program developed a plan when we first started using the standards and those get revisited and updated several times a year based on ongoing observation,”

explained Chrissy Poertner, who coordinates the accreditation and improvement process for the 18 programs. Observation data and program improvement plans are also used to guide staff development.

Initially some staff expressed concern that the tool was long and would be cumbersome to work with, but Poertner says the overall response has been very positive, especially because everyone is involved in and owns the process. “These tools give staff a guide, and when you’re out there working in the field the autonomy can feel overwhelming. Because we’ve created the buy-in and they are part of the improvement process, people respond really positively.”

For More Information

Additional information about NAA’s observation tool and accreditation process is available online at:
<http://naaweb.yourmembership.com/?page=NAACcreditation>

Contact

Judy Nee, President and CEO
The National AfterSchool Association
529 Main Street, Suite 214
Charlestown, MA 02129
888.801.3622
jnee@naaweb.org

Program Quality Observation Scale

Developed by Deborah Lowe Vandell & Kim Pierce

Purpose and History

The Program Quality Observation Scale (PQO), funded by the National Institute of Child Health and Human Development (NICHD) as part of an initiative to study out-of-school time, 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 tool has two components – qualitative ratings focused on the program environment and staff behavior and time samples of children's activities and their interactions with staff and peers.

The PQO was developed for research purposes by Deborah Vandell and Kim Pierce and has been used in a series of studies, primarily looking at the quality of school- and center-based after-school programs serving first through fifth grade elementary-school children. The instrument has its roots in Vandell's observational work in early childhood care settings, including the NICHD Study of Early Childhood Care and her work in after-school programs, including the Ecological Study of After-School Care funded by the Spencer Foundation.

The primary focus of the time sample procedure is on three components of individual children's experiences in programs – relationships with staff, relationships with peers and opportunities for engagement in activities. The qualitative ratings focus on all children's experiences in the program in terms of staff behavior and the program environment. The qualitative ratings of program environment are best suited for use in formal school- or center-based after-school programs, while the qualitative ratings of staff behaviors and the time sampling of children's activities and interactions are relevant in both formal program settings as well as informal, adult-supervised settings.

Content

The PQO was designed to help researchers understand the quality of children's experiences inside programs and focuses on three components of quality – relationships with staff, relationships with peers and opportunities for

engagement in activities. As noted above, the instrument has two major components – qualitative ratings and time samples of children's activities and interactions. Ratings are made of the program environment and staff behavior, or what the developers call "caregiver style." The following three aspects of the program environment are rated:

- Programming flexibility
- Appropriateness and diversity of the available activities
- Chaos

Four characteristics of caregiver style are rated:

- Positive behavior management
- Negative behavior management
- Positive regard for children
- Negative regard for children

The time sample component of the tool is designed to record the activities and interactions of individual children within the program. There are 19 different activity categories for observers to select from (e.g., arts/crafts, large motor, snack, academic/homework). In addition, the tool provides observers with six different types of interactions to look for: positive, neutral and negative interactions with peers, and positive, neutral and negative interactions with staff.

Because the focus of the PQO is on children's experiences inside of programs, it tends to focus primarily on social processes and less on resources or the organization of those resources within programs. However, Vandell and colleagues have developed a number of related measures that do capture aspects of these other components, such as a physical environment scale. Developed long before the National Research Council's features of positive developmental settings framework (2002), some aspects of the PQO align well with that framework while others more clearly reflect its early childhood roots.

Structure and Methodology

The first component of the PQO – the qualitative ratings – is focused on program environment and staff behavior or “caregiver style.” Ratings are assigned based on a minimum of 90 minutes of continuous observation. While program environment ratings are made of the program as a whole, caregiver style ratings are made separately for each staff member observed (but could be adapted to rate all staff members collectively).

Program environment and caregiver style ratings are made using a four-point scale. Users are given descriptions of what constitutes a 1, 2, 3 or 4 rating

<i>Chaos</i>	
4 =	Chaos and disorganization are highly characteristics, persisting across multiple activities and settings. The children are out of control. They may be fighting with one another, yelling, or behaving inappropriately, jumping on furniture, ruining materials, or just generally running around. Activities do not seem organized; disorder is evident.
3 =	There is chaos and disorganization in the environment, but it is not characteristic of many children or all activities. A group of children may exhibit the behaviors that merit a rating of 4 or some activities and transition times may be chaotic and disorganized such that the progress of or beginning of activities for some children is impeded.
2 =	One or two children’s behavior may be out of control, but in general, children’s behavior is appropriate and reasonably controlled. Transitions and activities generally go smoothly, although there may be exceptions.
1 =	No chaos or disorganization is observed in the environment. Children’s behavior is appropriate, and activities and transitions proceed smoothly.

for three distinct aspects of program environment – flexibility, activities and chaos, and four different aspects of caregiver style – positive behavior management, negative behavior management, positive regard and negative regard. A “4” rating means that particular aspect of the environment (or staff behavior) is highly characteristic of the program (see example below).

The time sampling component of the PQO is focused on the activities and interactions that individual children engage in at an after-school program. Activity type is recorded using 19 different categories. Interactions are assessed in terms of whether they are positive, neutral or negative and whether they happen with peers or with staff. In addition, staff interactions are further coded to note whether they are one-on-one, small group or large group.

Time sampling entails documenting the activities and interactions that a number of individual children have in a program for short periods of time. The developers of the PQO suggest that 30-minute time samples be conducted in 30-second intervals (for a total of 60 intervals). During each interval, the rater observes a child for 20 seconds and then spends 10 seconds recording or coding what they observed. Because time sample observations will sometimes involve fewer than 60 intervals, scores need to be adjusted for the total number of intervals actually observed. This time sampling component has been adjusted for use in different studies (for example with longer observation periods, fewer cycles, etc).¹³

Technical Properties

Available psychometric evidence supporting the PQO addresses score distribution, interrater reliability, test-retest reliability, convergent validity and concurrent validity, mostly from a report by Vandell and Pierce

¹³ Readers should note that the developers did not design the PQO for self-assessment, but rather for a large study that required that time sample observations center on a single child of interest. The time sample component of the instrument could be modified for general use by observers randomly choosing children for each assessment. However, it is unclear if the available psychometric findings on the time sample observations will extend to this modified instrument. This caveat does not apply to the qualitative ratings, which were designed to measure the program as a whole and do not require modification for self-assessment.

(2006) based on multiple observations of after-school programs over several years in the NICHD Study of After-School Care and Children's Development. The study included a broad sample of 46 for- and nonprofit programs located in schools, child care centers and community centers. Each program was observed three or four times a year over a five-year period. Predictive validity evidence comes from a study conducted by Pierce, Bolt and Vandell (2008), which examines social and academic outcomes from 120 children enrolled in 46 after-school programs. Children were assessed during their 2nd and 3rd grades.

Score Distributions

Score distributions help users determine whether items adequately distinguish between programs on specific dimensions. Vandell and Pierce examined the average scores and ranges for overall observed quality and the individual qualitative rating scales obtained in formal programs in the Study of After-School Care. The overall quality score was created by averaging the individual qualitative ratings (after reversing the scores for Chaos, Negative Regard and Negative Behavior Management). For both the overall quality score and the individual ratings, annual composites are averages of all observations conducted within a school year.

The overall score and most of the qualitative ratings and time-sampled activities and interactions had wide variability, suggesting the instrument can detect differences among a variety of programs. Across multiple observations in several years of study, the full range of scores was obtained for most of the sampled activities. Among the qualitative ratings, low variability was found among Negative Behavior Management and Negative Regard for Children. However, the strong validity evidence suggests that the instrument is detecting meaningful differences in these domains despite their low frequencies.

For children's interactions, the full range of scores was obtained for neutral interactions with staff and with peers; as would be expected, the range was more restricted for interactions that were clearly positive or negative.

Interrater Reliability

The degree to which different raters agree when observing the same program was tested for both the qualitative ratings and time sampling components of the instrument. For the qualitative ratings, kappa coefficients were computed once a year over four years. All of the domains had scores above .70, the benchmark for strong interrater reliability, except Staff Negative Regard, for which the lowest coefficient was .59. The proportion of Negative Regard scores on which observers achieved exact agreement was high, however, suggesting that the moderate kappa score may be due to the relative infrequency with which negative regard was observed. The average kappa score for staff negative regard was .82, suggesting that trained raters can reach acceptable agreement on all domains.

Agreement was also computed for all domains of the time sample observations except group size. Kappa scores at all time points were either above .70 or very close, indicating strong interrater reliability.

Internal Consistency

To determine whether items fit together to form a meaningful overall score, the authors computed a statistic called Cronbach's alpha. Vandell and Pierce found alpha levels for the annual overall observed quality scores averaged .81, well above the recommended cutoff of .70.

Test-Retest Reliability

In order to determine whether the quality composite and individual qualitative ratings generated by the PQO are stable over time, Vandell and Pierce correlated the ratings made in adjacent observations during the second year of the study, when the sample was largest and most representative of the range of programs in the community (N=45). Four observations were conducted in each program, approximately two months apart. Ratings from the first observation were correlated with those from the second; the ratings from the second observation were correlated with those from the third; and the ratings from the third observation were correlated with those from the fourth. Correlations

for overall observed quality were near or above .70, suggesting the instrument detects changes in program quality and is not overly sensitive to minor changes. Correlations for the individual ratings were lower, with the average for all domains ranging from .34 to .59. This suggests that programs are only somewhat stable in their scores for particular domains over periods of two months. It is unclear whether this reflects short-term variability (as might be seen from one day to the next) or meaningful changes over the course of two months.

Convergent Validity

To examine whether the PQQ yields accurate information about the aspects of programs it is supposed to measure the authors compared findings for the qualitative ratings to findings from the SACERS (also reviewed in this report). If both instruments truly measure program quality, one can reasonably expect that the findings will be related. The following relationships between PQQ and SACERS scales were examined: (1) PQQ Programmatic Flexibility was positively related to SACERS Program Structure, (2) PQQ Available Activities was positive related to SACERS Activities, and (3) PQQ Staff Positive Regard and Positive Behavior Management was positively related to SACERS Interactions, and PQQ Staff Negative Regard and Negative Behavior Management was negatively related to SACERS Interactions. These findings provide strong evidence that the instrument adequately measures program quality. Although convergent validity is supported for most qualitative items, we cannot infer the validity of the Chaos rating because there was no comparable SACERS question to compare it to.

Concurrent Validity

Another way to examine whether the PQQ yields accurate information about the aspects of programs it is supposed to measure is to compare the instrument's ratings to other distinct but theoretically important concepts. Developers compared the findings assessed by the PQQ to structural features of after-school programs (Pierce, K. M., Hamm, J. V., Sisco, C., & Gmeinder, K., 1995). Similar to what is reported in the early childhood literature, Positive Regard ratings were higher and Negative Regard scores were lower in nonprofit

programs compared to for-profit programs, when child-staff ratios were smaller and when program staff had more formal education. Programming flexibility was higher in nonprofit compared to for-profit programs and when child-staff ratios were smaller. Ratings of available activities were higher in nonprofit programs.

Relations between time-sampled activities and interactions were also associated with program characteristics as well as child reports of their experiences in the programs. For example, children were observed to have more frequent positive/neutral interactions with staff and less frequent negative interactions with peers in programs with smaller group sizes; and smaller staff-child ratios were associated with children having more frequent positive/neutral interactions with staff and spending less time in transition (e.g., standing in line) and in large motor activities. Although this evidence is quite strong, it is unclear whether researchers had expected additional relationships that they did not find.

Predictive Validity

In a study conducted by Pierce, Bolt, and Vandell (2008), researchers examined the relationships between three PQQ scales (Staff-child relations, Available activities, and Programming flexibility) and social and academic outcomes of 120 children enrolled in 46 after-school programs. Program observations were conducted several times per school year for two years, and children's outcomes were assessed at the end of each school year when they were in 2nd and 3rd grades, respectively. Better Staff-Child relations was associated with higher reading scores (both grade levels), math scores (grade 3 only) and better social skills (2nd grade boys only), but the scale was unrelated to work habits. Activities was related to better math grades and work habits (both for Grade 3 only), but the scale was unrelated to reading grades and social skills. Programming Flexibility was not related to any outcomes. Predictive validity appears mixed but the evidence should be regarded as preliminary. The authors state that little research currently exists that examines the relationships of specific program strategies (rather than overall program quality) with children's academic and social outcomes.

User Considerations

Ease of Use

While the PQO is available for anyone to use, it is important to recognize that it was developed with exclusively a research audience in mind. While the manual includes basic instructions for conducting observations and completing the forms, it was written for researchers participating in data collection related to a particular study. The materials have not been tailored for general or practitioner use at this time and therefore include some concepts and language (e.g., adjusted frequencies, sampling, qualitative) that may not be particularly accessible for non-research audiences.

In the context of the studies the PQO was developed for, formal observation time at sites was fairly limited, but some additional time should be factored in for reviewing notes and assembling ratings. It is recommended that the qualitative ratings of environment and staff behavior be made based on a minimum of 90 minutes of observation. Completing the time sample process as outlined in the manual takes a minimum of 30 minutes (60 30-second cycles) for an experienced observer. Some guidance about how to conduct observations and develop ratings is provided in the manual.

Available Supports

At this time, training is not regularly available on how to use the PQO, but has been conducted with data collectors involved in the studies it was developed for. Trainings include reviewing the contents of the instrument and pairing new raters with trained raters to do an observation in the field, compare scores and build inter-observer agreement.

Observation data collected using the PQO have always been coupled with supplementary data sources such as a questionnaire about the physical environment as well as staff, student and parent surveys. However formal links do not exist between the observation tool and other measures and the PQO could be used independently.

In the Field

In the Study of After-School Care and Children's Development, conducted by Deborah Vandell and Kim Pierce in the mid-1990s, live observation of children's experiences in programs was at the center of the research (Pierce, K., Hamm, J., & Vandell, D.L., 1999). Observations were conducted during the program participants' first-grade year and each child was observed three times by an individual observer who was randomly assigned from a pool of observers. The observers used both components of the PQO – the time sample procedure and qualitative ratings of the program environment and caregiver style. Other types of information were collected using different methods and measures.

In analyzing the data, the researchers looked for associations between the various measures of program quality and also at associations between program quality and children's adjustment at school. In terms of how aspects of program quality relate, staff positivity was negatively correlated with staff negativity, as one might expect. Staff positivity was higher in programs that were more flexible and offered more activities. Staff negativity was associated with less programmatic flexibility. They also found associations between the program quality indicators and children's adjustment in the first-grade classroom, primarily for boys. Staff positivity was associated with boys earning higher reading and math grades and exhibiting less externalizing behavior at school. Greater programming flexibility was associated with boys exhibiting better social skills at school. Greater availability of age-appropriate activities was associated with boys earning poorer reading and math grades, and exhibiting poorer work habits and more externalizing behavior at school.

Pierce, Bolt, and Vandell (in press) recently examined associations between program quality indicators as measured by several PQO qualitative ratings (staff positive regard, activities, flexibility) and children's adjustment at school (grades, work habits, social skills) in Grades 2 and 3, controlling for child and family characteristics and child prior adjustment. The

researchers found that greater staff positivity in after-school programs was associated with both boys and girls earning better reading grades in Grades 2 and 3 and better math grades in Grade 3. Boys also exhibited better social skills in Grade 2 when their after-school programs were characterized by greater staff positivity. Availability of multiple activities in the after-school programs was associated with boys and girls earning better math grades and exhibiting better work habits at school in Grade 3. Programming flexibility was not associated with child outcomes in Grades 2 and 3.

Vandell and Pierce (2001) also reported long-term associations between overall program quality, as measured by annual composites of the qualitative ratings and children's outcomes. They looked at cumulative program quality (averaged across two years, three years and four years) in relation to children's adjustment at school. Controlling for child and family characteristics and for children's functioning at the end of first grade, the researchers found that children who experienced greater cumulative program quality in Grades 1–3 were reported by their teachers to have better academic grades at school. Girls whose after-school programs had higher cumulative quality across Grades 1–3 or 1–4 had better work habits and better social skills with peers at school in Grades 3 and 4.

For More Information

The PQO is available online at:
<http://childcare.gse.uci.edu/des4.html>

Contact

Deborah Lowe Vandell
Department of Education
University of California, Irvine
2001 Berkeley Place
Irvine, CA 92697
949.824.7840

Program Quality Self-Assessment Tool

Developed by the New York State Afterschool Network

Purpose and History

In 2003, the New York State Afterschool Network (NYSAN) began a two-year process of developing the Program Quality Self-Assessment Tool (QSA). A Quality Assurance committee involving key stakeholders from practice, policy and research, reviewed relevant literature, drafted the instrument, conducted field tests and incorporated feedback from practitioners across the state. Soon after the instrument was completed in 2005, New York State began requiring that all 21st CCLC-funded programs use it twice a year for self-assessment purposes.

The QSA was developed exclusively for self-assessment purposes; programs are discouraged from using it for external assessment or formal evaluation. It is intended to be used in its entirety, ideally as the focal point of a collective self-assessment process that involves all program staff. The QSA is also used by new after-school programs during their initial development; specific items that are considered “foundational” indicators for the start-up stage are identified.

The QSA was designed to be used in the full range school and community-based after-school programs and 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). It was also designed to be used by programs serving a broad range of students, from kindergarten through high school.

Content

The Program Quality Self-Assessment Tool is organized into 10 essential elements of effective after-school programs (see below). Each element contains a list of standards of practice or quality indicators that describe each element in greater detail. The elements represent a mix of activity-level, program-level and organizational-level concerns:

- Environment/Climate
- Administration/Organization

- Relationships
- Staffing/Professional Development
- Programming/Activities
- Linkages Between Day and After-School
- Youth Participation/Engagement
- Parent/Family/Community Partnerships
- Program Sustainability/Growth
- Measuring Outcomes/Evaluation

Because the QSA was designed with an eye towards programs receiving 21st CCLC funding, there was an intentional effort to capture aspects of programming that although they may not relate directly to academics, will enhance programs’ ability to address students’ educational needs. The developers are exploring options that would allow programs to address a subset of items based on their level of readiness; however, the ultimate goal is to assess the program or organization in its entirety.

Because of its broad focus extending from the activity level to the organization as a whole, the QSA emphasizes several different components of program settings including social processes, program resources and the organization or arrangement of those resources inside the program. Social processes addressed by the tool include relationships, climate and pedagogy. Resource issues include facilities and staffing requirements and arrangements such as effective transitions, policies and procedures and relationships with schools are also addressed.

Structure and Methodology

Because of its commitment to child and youth development broadly defined, it is not surprising that the items included in the QSA reflect each of the features identified by the National Research Council as features of positive developmental settings (2002).

Each of the QSA’s 10 essential elements of effective after-school programming is further defined by a

<i>Relationships: A QUALITY program develops, nurtures and maintains positive relationships and interactions among staff, participants, families and communities.</i>							
<i>A Quality Program:</i>	<i>Performance Level</i>				<i>Plan to Improve</i>		
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>Right Now</i>	<i>This Year</i>	<i>Next Year</i>
<i>Has staff who respect and communicate with one another and are role models of positive adult relationships.</i>							
<i>Interacts with families in comfortable, respectful and welcoming way.</i>							
<i>Treats participants with respect and listens to what they say.</i>							
<i>Teaches participants to interact with one another in positive ways.</i>							
<i>Teaches participants to make responsible choices and encourages positive outcomes.</i>							
<i>Is sensitive to the culture and language of the participants.</i>							
<i>Establishes meaningful community collaboration.</i>							
<i>Has scheduled meetings with its major stakeholders.</i>							
<i>Encourages former participants to contribute as volunteers or staff.</i>							

summary statement, which is then followed by between 7 and 18 quality indicators – statements aimed at illustrating what a particular element looks like in practice. 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.

The rating scale used in the QSA (see example below) 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. Staff are asked to determine whether their performance in each indicator area is:

- 4 = Excellent/Exceed Standards
- 3 = Satisfactory/Meets Standards
- 2 = Some Progress Made/Approaching Standard
- 1 = Must Address & Improve/Standard Not Met

While some additional guidance is provided to staff in the tool's introduction about how to determine ratings, developers acknowledge that this is one of the areas they may revisit in the future, based on feedback from the field. Users are not encouraged to combine scores for each element or to determine a global rating, because the tool is intended for internal self-assessment purposes only. In addition to assigning a rating for each indicator, users are given space on the form to note and prioritize their plans for improvement.

Technical Properties

Beyond establishing face validity (people with expertise in the after-school field agree this measures important features of program quality), research related to the instrument's psychometric properties has not been conducted.

User Considerations

Ease of Use

Practitioners led the development of the QSA and represent its primary target audience. The language and format of the instrument are straightforward and user-friendly. The tool consists of one document, free and downloadable from the Web, that includes an overview, instructions and the instrument itself.

NYSAN has developed a new user guide, published in April 2008, to assist programs in utilizing the QSA. This tool provides guidance on how to engage staff in the assessment process in addition to outlining the basic guidelines for administering the tool.

Programs are expected to go through the self-assessment process twice a year. Some in the field have concerns about the tool being too lengthy; this feedback will be taken into an upcoming revision process.

Additional Supports

The user guide, mentioned above, was created in consultation with a wide range of stakeholders – including NYSAN staff, a state-wide Quality Assurance Committee, practitioner-based focus groups and an advisory group. The guide serves as a “self-guided

walkthrough” the QSA tool; the tool is embedded in the second half of the guide. NYSAN is currently developing phase two of the guide – an online version which will allow users to click on links to other web-based tools, articles and resources related to anyone of the ten essential elements or the overall quality assessment and improvement process. The online version will also provide a descriptive example of optimal performance for every single indicator (the current hardcopy guide features only select examples). Programs can contact NYSAN to receive additional referrals for technical assistance in using the instrument.

While no centralized mechanism for collecting or analyzing results currently exists, with the development of the online version of the tool and user guide, it will be possible to enter data by computer. This could lead to efficient opportunities to track and analyze data over time.

Although additional instruments are not provided with the tool, users are encouraged to consider QSA results one important source of data to inform program planning and are encouraged to use it in concert with other formal or informal evaluative efforts such as participant, parent and staff surveys, staff meetings and community forums. In the future, users will be able to link to other tools from the online version of the QSA and guide.

All NYSAN training is now organized by the 10 elements featured in the tool, so practitioners can easily find professional development opportunities that connect with the results of their self-assessment. Regular trainings that are conducted twice a year with 21st CCLC grantees are also organized around the 10 elements.

In the Field

The Niagara Falls School District has funding through the 21st CCLC program to run after-school programs at four sites – three middle schools and one high school. While all after-school programs receiving 21st CCLC funding in the state of New York are required to conduct and submit QSA assessments twice a year, these programs in Niagara Falls have extended their use of the

tool well beyond self-assessment. They see the QSA as central to staff and program development efforts.

Susan Ross, the Program Director within the school district, described how site coordinators see the tool. “We see the QSA as a staff development resource. About three weeks after the school year starts, site coordinators begin sitting down with all of their staff – teachers and community partners – and walking through the tool, one page per staff meeting. It gives us a collective sense of what’s working and what we need to improve. It’s a great focal point for discussions among staff.”

Ross emphasized that one of the important benefits of this process is that it helps to level the playing field between staff from external community partner organizations and school teachers who work in after-school programs. “This really gives our partners an opportunity to feel their opinions are valued. Often when CBO staff come into schools they feel like guests as opposed to full-fledged partners. Through this process, they see their opinions are equally valued and that helps build overall staff morale.”

Site directors and staff find the tool accessible and user-friendly. Ross summed up her assessment of the QSA in a matter-of-fact way. “We like it. It’s easy to use, self-explanatory and understandable. In fact, I wouldn’t change anything about it.”

For More Information

NYSAN’s Program Quality Self-Assessment Tool is available online at:
www.nysan.org/content/document/detail/1991/

Contact

Ajay Khashu, Director of Research
NYSAN
925 Ninth Avenue
New York, NY 10019
212.547.6908
akhashu@nysan.org

Promising Practices Rating Scale

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

Purpose and History

The Promising Practices Rating Scale (PPRS) was developed for research purposes and is designed for use in school- and community-based after-school programs that serve elementary and middle school students. The tool allows observers to document the type of activity, the extent to which promising practices are implemented within activities and overall program quality.

The 2005 version of the PPRS, the version that is currently available, was developed by Deborah Vandell, Liz Reisner, Kim Dadisman, Kim Pierce and Ellen Pechman in the context of a specific study focused on the relationship between participation in “typically performing” programs and child and youth outcomes (Vandell, D., Reisner, E., Pierce, K., Brown, B., Lee, D., Bolt, D., Dadisman, K. & Pechman, E., 2006; Vandell, D., Reisner, E. & Pierce, K., 2007). Because of this, the tool was initially designed to verify whether or not programs were high-quality rather than to look at variations in quality across programs.

This instrument builds directly on earlier work by Vandell and colleagues focused at the elementary level (see write up of the Program Quality Observation Scale in this report) as well as the features of positive developmental settings identified by the National Research Council (2002). Its authors also drew upon several other observation instruments included in this report as they developed the exemplars of promising practices: the School-Age Care Environment Rating Scale, the Program Observation Tool and the OST Observation Tool designed by Policy Studies Associates.

Although the focus of this summary is the PPRS specifically, other components of the Promising Practices quality assessment system include interviews and questionnaires completed by program directors and staff. These tools obtain information about structural features of programs such as staff qualifications and ongoing training, material and financial resources and connections between the program and school, family and community.

Content

The PPRS provides researchers with a framework for observing essential indicators of high quality programs. It addresses three different aspects of programming: activity type, implementation of promising practices and overall program quality. The first section, which closely mirrors the OST Observation Tool developed by Policy Studies Associates, focuses on documenting a range of in-depth information about the type of activity being observed and the skills emphasized through that activity. The promising practices ratings that constitute the core of the instrument focus on the following eight areas of quality:

- Supportive Relations with Adults
- Supportive Relations with Peers
- Level of Engagement
- Opportunities for Cognitive Growth
- Appropriate Structure
- Over-control
- Chaos
- Mastery Orientation

Because of its emphasis on what children and youth experience in programs, the PPRS has an activity and program-level focus and does not address organizational issues related to management, leadership or policy. The primary focus is on social processes – including interactions between and among youth and staff and some aspects of instruction.

As mentioned above, the developers drew heavily on the *Community Programs to Promote Youth Development* report (National Research Council, 2002), so the features of positive development outlined in that report are quite visible within the tool’s definition of promising practices. Although the PPRS itself does not include a focus on connections between the program and school, family or community (one of the features described in the NRC report), companion tools are available to capture this type of information.

Structure and Methodology

The first part of the PPRS, which focuses on the activity context, has observers watch an activity for 15 minutes and code several aspects of what they are observing, including:

- Activity type (e.g., tutoring, visual arts, music, sports, community service);
- Space (e.g., classroom, gym, library, cafeteria, auditorium, hallway, playground);
- Primary skill targeted (e.g., artistic, physical, literacy, numeracy, interpersonal);
- Number of staff involved in the activity; and
- Number, gender and grade level of participants.

These observations are recorded on a coversheet that also includes other basic information about the observer, the program, date, time, etc.

Next observers are asked to write down a brief narrative description of the activity they are observing, following a set of specific guiding questions (see below). This description supplements the activity context coding with a richer description of what is going on.

- What are youth doing?
- What kinds of materials are used?
- What kinds of instructional processes are used?
- What, if any, specific skills does the activity’s leader(s) have that supports the instruction involved in the activity he/she is conducting?

Level of Engagement (in intended experiences)	
<i>High</i>	<i>Low</i>
<p><i>Students appear engaged, focused and interested in their activities.</i></p> <ul style="list-style-type: none"> • Engaged in the focal activity and/or using free time appropriately. • Appear to be interested in the activity. • Follow staff directions in an agreeable manner. 	<p><i>Students appear bored or distracted.</i></p> <ul style="list-style-type: none"> • Ignore staff who are talking to them • ‘Pretend’ to listen. • Wander aimlessly.
<p><i>Markers of engagement are appropriate to activity (e.g. intense concentration witnessed during computer activity, high levels of affect during sports activities; can be solitary or group activities).</i></p>	<p><i>Markers of engagement inappropriate to activity (e.g. picking flowers while playing a sport activity).</i></p>
<p><i>Students contribute to discussions.</i></p> <ul style="list-style-type: none"> • Discuss back and forth and offer comments. • Ask “on-task” questions. • are comfortable initiating conversation. 	<p><i>Students do not contribute to discussions.</i></p> <ul style="list-style-type: none"> • Do not participate in discussions. • Do not ask questions.
<p><i>Rating Indicators:</i></p> <p>1 = Most students are not engaged appropriately, may appear bored.</p> <p>2 = Students are participating in activities but do not appear to be concentrating or affectively involved.</p> <p>3 = Students are focused on activities with some evidence of affective involvement or sustained concentration.</p> <p>4 = Students are concentrating on activities, focused, interacting pleasantly when appropriate and are affectively involved in the activity.</p>	

- What is the overall affective tone?
- To what extent are youth engaged?
- Describe observed promising practices as appropriate and raise concerns about quality, if there are any.

The next section and the core of the PPRS is the Promising Practices Ratings section, where observers document to what extent certain exemplars of practice are present in the program. This section of the tool addresses the eight key areas of practice listed previously.

Each area of practice is subdivided into two to five specific exemplars, with more detailed indicators provided under each. Observers are given both positive and negative exemplars and indicators for each practice area in order to help guide determination of ratings (see example on page 63).

In the PPRS, ratings are only assigned at the overall practice level (not for individual exemplars or indicators). Practices are either considered highly characteristic (4), somewhat characteristic (3), somewhat uncharacteristic (2), or highly uncharacteristic (1). Additional guidance as to what each of these terms mean is provided in the instrument. At the bottom of the description of each practice area, observers are given tailored guidance as to what might lead to a 1, 2, 3 or 4 rating for that practice.

Finally, observers are asked to review their ratings of promising practices across multiple activities and assign an overall rating for each promising practice area. An overall program quality score is computed as the mean of the ratings on the 8 scales, after reversing the scores for over-control and chaos. For each practice area there is space to write down notes to “justify” the overall rating assigned.

Technical Properties

Available psychometric evidence supporting the PPRS addresses interrater reliability, score distribution, internal consistency and predictive validity information from a study of 35 after-school elementary and middle school programs (Vandell, D., Resiner, E., et.al., 2006).

Score Distributions

Score distributions help users determine whether items adequately distinguish between programs on specific dimensions. Vandell, Reisner, et al. (2006) examined the average scores for overall program quality and the individual rating scales obtained with the sample of high-quality elementary and middle school programs that participated in the Study of Promising After-School Programs sample at two time points. Generally, it is important to have a range of scores across programs, as that would suggest the measure can detect meaningful differences between programs. Because this sample included only high-quality programs, however, the scores naturally fell toward the positive extremes of each dimension. Score distributions on the PPRS obtained in 37 observations in programs of varying quality are more widely distributed, suggesting that the instrument detects meaningful differences among programs.

In that study, the authors theorized that scores would exhibit a wider range and would show low, moderate and high quality (as opposed to most programs scoring on the high end of the scale). As expected, scores for each of the scales generally had a wider distribution with averages across the programs falling in the middle for most of the eight scales. Two scales, Opportunities for Cognitive Growth and Mastery Orientation, had low scores overall (averages were 1.65 and 1.78 on a scale of 1 to 4, respectively), although their scores in the Study of Promising Practices were closer to the center (average scores across two observation periods within programs serving elementary and middle school students, respectively, ranged from 2.6 to 2.9). The distribution differences for these two scales suggest that they may be better suited to differentiate among programs of higher quality. The Overcontrol scale was the only item that was consistently low for both the Study of Promising Practices and the follow up study, which may simply suggest that staff in most programs do not exhibit a great deal of overcontrol. Taken together, the two studies provide strong evidence that the instrument captures meaningful differences across a variety of programs.

Interrater Reliability

The authors examined rater agreement for each of the instrument's eight items by calculating intraclass correlation coefficients between ratings of 24 programs made by two observers. Coefficients ranged from .58 for Opportunities for Cognitive Growth to .86 for Structure (average=.74) for the individual scales. The intraclass correlation for the overall program quality score was .90. Interrater agreement represented by kappa scores in work conducted by other research teams in programs of varying quality ranged from .63 for over-control to .94 for supportive relations with adults (average=.77) across 37 observations made by two observers. These scores indicate acceptable interrater reliability, meaning the instrument's items are clear enough for raters to understand and agree on.

There is currently no information regarding interrater reliability for the overall program quality rating, which range on a three-point scale from low program quality to high program quality. Therefore, evidence does not exist to support whether raters agree on the degree of quality that individual programs exhibit.

Additional Reliability Evidence

Additional rater agreement information was obtained by comparing two sets of ratings by the same rater conducted on consecutive days for each program. The authors found that the percent agreement for ratings of each feature over two days was between 81 percent and 97 percent, with an average of 90 percent. This translates into an average kappa score of 0.80, indicating that the average item's rating for Day 2 is not very different from Day 1.

Internal Consistency

To determine whether the items fit together to form a meaningful overall score, the authors computed a statistic called Cronbach's alpha. In the Study of Promising After-School Programs, alpha coefficients for the overall program quality score ranged from .74 to .77, indicating acceptable internal consistency.

¹⁴ Results were found using two advanced statistical techniques known as cluster analysis and hierarchical linear modeling.

Predictive Validity

Initial evidence of predictive validity is available for the PPRS, which means that the instrument does predict youth outcomes that would be expected from prior theory or research. Specifically, Vandell, Reisner and colleagues (2005) found that youth attending high quality programs (as measured by the PPRS) had better educational and behavioral outcomes by the end of the academic year than unsupervised youth who did not regularly attend any after-school program, including better academic performance, task persistence, social skills and pro-social behaviors with peers and less misconduct, substance abuse and aggressive behavior.¹⁴ Vandell, Reisner et al. (2006) reported similar findings for longer term outcomes after two years of program participation. Improvement in math achievement scores have also been reported (Vandell, Reisner & Pierce, 2007).

The fact that the instrument's ratings related to expected outcomes offers some reassurance to users that it accurately measures aspects of program quality. However, the validity evidence should be taken as preliminary for several reasons. First, the authors have not examined PPRS ratings of low-quality programs. No evidence exists that the instrument distinguishes between expert ratings of low- and high-quality programs, or whether low-quality program ratings predict youth outcomes differently than high quality program ratings. It would also be useful to understand the predictive validity of each specific scale (e.g., level of engagement, appropriate structure) and the overall score.

User Considerations

Ease of Use

While the PPRS is available online and free for anyone to download and use, it is important to recognize that it was developed with primarily a research audience in mind. While the observation manual includes basic instructions for conducting observations and completing the forms, it was written for researchers participating in data collection related to a particular study. The materials have not been tailored for general use or for practitioner use at this time and therefore include some language (e.g., construct, exemplar) that may not necessarily be accessible for non-research audiences.

In the context of the study the PPRS was developed for, site visits were fairly time-intensive (spread over the course of two days). However, formal observation time totaled approximately two hours per site, with several additional hours spent reviewing notes and assigning ratings. Some additional guidance about how to conduct observations, develop ratings and complete the forms is provided in the manual.

Available Supports

At this time, training is not regularly available on how to use the PPRS, but has been conducted with data collectors involved in research. Trainings have included reviewing the contents of the instrument and pairing new raters with trained raters to do an observation in the field, compare scores and build inter-observer agreement.

Observation data collected with the PPRS has always been coupled with supplementary data sources such as a questionnaire about the physical environment as well as staff, student and parent surveys. However formal links do not exist between the observation tool and other measures and the PPRS could be used independently. Additional measures are also available at the same web site as the PPRS.

In the Field

The Promising Practices Rating System was developed specifically for use in the Study of Promising After-school Programs, a national study funded by the C.S. Mott Foundation that focused on the short- and long-term impacts of high-quality after-school programs on the cognitive, academic, social and emotional development of children and youth in high-poverty communities. The research was led by Deborah Vandell of UC Irvine (formerly of the University of Wisconsin-Madison) and Elizabeth Reisner of Policy Studies Associates.

Two-day site visits to participating programs were conducted in fall 2002, spring 2003, fall 2003 and spring 2005 to assess the quality of each program. During sites visits, researchers conducted observations using the PPRS on two afternoons, for a minimum of one hour per day. Observers focused on the activities of

the target age groups (grades 3 and 4 and grades 6 and 7) and observed as many different types of activities as possible, with a minimum of 15 minutes per activity. At the end of the first day of the site visit, observers assigned tentative ratings to each of the eight practice areas; at the end of the second day, the final ratings were determined.

As analyses got underway, the authors revised their conceptual scheme based on the idea that sets of experiences should be taken into consideration as opposed to labeling students as program vs. non-program. Elementary students with high rates of participation in quality after-school programs but low levels of participation in other after-school arrangements (the program only cluster) outperformed the low supervisor cluster (self-care + limited activities) on every measure of academic and social competence assessed. The supervised at home group outpaced the self-care + activities cluster on all academic measures and social skills. Among middle school students, the program+ activities cluster had better work habits than the low supervision group and both program groups (program + activities, program only) reported less misconduct and substance use compared to the low supervision group. Similar results were found for program involvement across three years. Additional findings are available at <http://childcare.gse.uci.edu/des3.html>.

For More Information

The PPRS is available online at:
<http://childcare.gse.uci.edu/des3.html>

Contact

Deborah Lowe Vandell
Department of Education
University of California, Irvine
2001 Berkeley Place
Irvine, CA 92697
949.824.7840
dvandell@uci.edu

Quality Assurance System[®]

Developed by Foundations, Inc.

Purpose and History

The Quality Assurance System[®] was developed by Foundations, Inc. to help after-school programs conduct quality assessment and continuous improvement planning. Based on Foundations, Inc.'s experience running after-school programs, offering professional development activities and providing technical assistance and publications for the field, the QAS was designed to help programs develop and sustain a commitment to quality.

In its first incarnation, the QAS was a simple checklist designed to assess the quality of after-school programs operated by the organization itself. Roughly five years ago staff at Foundations reconstructed and expanded the tool for broader use, with input from practitioners both inside and outside of the organization. The QAS was developed to be general enough for use in a range of school- and community-based programs serving children and youth grades pre-K–12.

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 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. Filling out the QAS requires a combination of observation, interview and document review. Scores are generated for each building block rather than the overall program, reflecting the tool’s emphasis on identifying specific areas for improvement.

Content

The various components of quality that the QAS focuses on are called “building blocks.” The seven core building blocks, which describe what Foundations considers to be the fundamental features that underlie effective after-school programming, include:

- Program planning and improvement;
- Leadership;
- Facility and program space;

- Health and safety;
- Staffing;
- Family and community connections; and
- Social climate.

In addition to these seven, three “program focus building blocks” reflecting the particular goals or focus of a program are available for users to select from:

- Academics;
- Recreation; and
- Youth development.

The QAS puts roughly equal emphasis on three different components of settings including social processes, program resources and the arrangement or organization of those resources within programs. There are items on the QAS that address all of the features of positive developmental settings outlined by the National Research Council (2002), with somewhat more of an emphasis on things related to structure and skill-building than on features such as “support for efficacy and mattering” and “supportive relationships.”

Structure and Methodology

The structure of the QAS is clear and straightforward. Part one – program basics – includes the seven core building blocks. For each one, users are given a brief description of the importance of that aspect of quality. The building block is further subdivided into five to eight specific elements, each of which are assigned a rating by assessors. For example, the elements of the social climate building block include: behavioral expectations, staff/participant interactions, diversity, social time and environment. For each element, more specific descriptions (also referred to as a “rubric”) are provided. Part two of the tool – program focus – consists of the three additional building blocks and its structure parallels that of part one. Programs are encouraged to use one, two or all three of the program focus building blocks in conducting their assessment.

Ratings for the QAS are made using a four-point scale from unsatisfactory (1) to outstanding (4). For each element of a building block, specific descriptions of what might lead to a 1, 2, 3 or 4 rating are provided (see example below).

In terms of data collection, users are provided with a document checklist that identifies what kinds of specific documents might be useful in filling out the QAS and are encouraged to gather and examine such documents prior to observing the program. The “program profile”

section of the tool asks users to upload important basic information about the program and can also be filled out, for the most part, prior to visiting.

- Once on-site, the users’ guide encourages observers to go through five steps:
- Meet people to establish rapport and hear from staff and youth about the program;
- Wander with purpose to develop a sense of the entire facility;

<i>Staffing</i>	<i>Unsatisfactory</i>	<i>Needs Improvement</i>	<i>Satisfactory</i>	<i>Outstanding</i>	<i>Score</i>
<i>Elements</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	
5.1 Staff to Participant Ratio	Insufficient staff are hired for the number of participants	Sufficient staff are hired for some levels of participation, but staffing is sometimes insufficient due to attendance fluctuations	Appropriate participant to staff ratios are maintained consistently.	Staff number and attendance exceed required ratios.	
5.2 Qualifications	Fewer than half the staff have the required training and/or experience	More than half the staff have the required training and/or experience.	All staff have the training and/or experience required by the program.	Many staff members exceed training and/or experience required by the program.	
5.3 Professional Growth	Professional development is not provided nor is time allocated for staff to pursue individual professional growth.	Some professional development opportunities are provided, but they are poorly attended.	Staff attend professional development sessions at least twice a year.	Staff identify professional development needs and attend professional development sessions more than twice a year.	
5.4 Attendance	Staff absenteeism is an ongoing problem (e.g. significant number of staff routinely absent).	Staff absences are an occasional problem.	Staff are reliable and absences are infrequent.	Staff absences are rare.	
5.5 Retention	Staff turnover is identified as a problem.	Staff turnover occasionally affects program offerings.	Staff retention is not identified as a problem.	Staff retention is excellent and provides stability.	
Total Score					

- Observe activities to see the program in action, the level of engagement and the nature of activities;
- Gather materials to ensure that all of the documents in the checklist and any other relevant materials are collected; and
- Take notes to ensure you have a running record of your observations and questions.

Once scores for each element are entered into the QAS, the program electronically generates overall building block scores. The program's quality profile then begins to emerge through summary graphs the software generates for each building block as well as a program summary graph that contains scores for each building block assessed. The graphs and building block scores help users target areas for improvement as part of the assessment process. A follow-up QAS assessment enables users to identify areas of progress and then refine goal-setting and improvement planning.

Technical Properties

Beyond establishing face validity (people with expertise in the after-school field agree this measures important features of program quality), research related to the instrument's psychometric properties has not yet been conducted.

User Considerations

Ease of Use

The QAS is a straightforward, flexible tool with several built-in features that make it particularly user-friendly. The instruction guide is written in clear, accessible language and walks users through the necessary background and basic steps for using the system. The standard cost for the QAS has recently been reduced to \$75 for an annual site license. This license is good for two official uses (or assessments) – which is what its developers suggest programs conduct annually, once toward the beginning of the year and once toward the end. After two uses the system generates a cumulative report comparing the initial and follow-up assessments. For programs with multiple sites, a cumulative report comparing site results is available with the initial

assessment. When the QAS is used as part of a professional development package related to quality improvement, discounts are available.

Available Supports

Foundations, Inc. offers online sessions and in-person training options to assist organizations in using the tool. Multi-site organizations may contract for individualized technical assistance and training, which may include options for customization of the tool. Trainings addressing quality elements reflected in the building blocks are available online, in technical assistance, and in professional development sessions.

For self-assessment purposes, once a QAS site license is purchased, programs can receive light phone technical assistance free of charge from Foundations, Inc. staff if they have questions while using the system. Programs that wish to have trained assessors conduct their assessment 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. The site-specific reports generated are specifically designed to help site staff and leaders use the information to guide improvement planning.

In the Field

Foundations, Inc. is working with the U.S. Dream Academy, a national program serving elementary and middle school students who are children of prisoners. In 10 centers around the country, U.S. Dream Academy provides a comprehensive program including academic support, enrichment, and a one-on-one mentoring relationship. The U.S. Dream Academy chose to use the QAS and a technical assistance strategy during 2008-2010 with Foundations to build and support program quality, and to establish an ongoing process of continuous improvement. During initial meetings, U.S. Dream leaders and staff worked with Foundations to clarify quality indicators for the program and customize the tool. The processes of co-assessment and self assessment are designed to build the capacity of sites

to target specific improvement goals and concrete steps, identify site strengths and innovations, and share strengths organization-wide. U.S. Dream Academy national headquarters will use the QAS findings to regularly identify where and how they can best support their centers.

Directors anticipate that the QAS and surrounding processes will direct staff to ask critical questions about their program environments and staff practice. At the same time, it will allow Centers to highlight and share strengths and accomplishments across the organization, building internal resources for quality. After joint assessments are conducted with U.S. Dream staff and Foundations at each site, individual scores will be aggregated and presented at a national meeting. Each center will conduct a follow-up self-assessment at the end of the 2008-2009 school year, at which point they will be able to analyze the data and evaluate their own program development. The QAS tool also will be available the following school year to allow each site to continue the self-assessment process.

The QAS design, coupled with technical assistance processes, allows for customization of the tool. With the addition of an eleventh building block focused on U.S. Dream Academy's mentoring component, the tool encompasses the organization's full range of essential program components. C. Diane Wallace Booker, Executive Director of the U.S. Dream Academy, stated that, "the beauty of QAS is that it is designed with evidence-based quality indicators yet is customizable and able to capture the unique elements of our program and helped us to more clearly define what quality looks like for us versus any other afterschool program. Further, the process of guided, self assessment and continuous improvement planning is critical to our ongoing efforts to achieve impact in the lives of the children we serve." Establishing an ongoing process for quality-building tailored to the specifics of the program is particularly important for multi-site programs. As U.S. Dream Academy expands to serve more children, the sustained quality assurance component becomes ever more critical.

For More Information

Additional information about the QAS, including ordering information is available online at:
<http://qas.foundationinc.org/start.asp?st=1>
 or by visiting www.afterschooled.org

Contact

Rhe McLaughlin, Associate Executive Director
 Center for Afterschool Education
 Foundations, Inc.
 Moorestown West Corporate Center
 2 Executive Drive, Suite 4
 Moorestown, NJ 08057
 856.533.1600
rmclaughlin@foundationinc.org

School-Age Care Environment Rating Scale

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

Purpose and History

The School-Age Care Environment Rating Scale (SACERS) is designed to assess before- and after-school care programs for elementary school age children (5- to 12-year olds) as well as whole day programs in communities with year-round schools. It focuses on “process quality” or social or educational interactions going on in the setting, as well as program features related to space, schedule, materials and activities that support those interactions.

The SACERS was developed for self-assessment, program monitoring or program improvement planning, as well as for research and evaluation. It can be used by program staff as well as trained external observers or researchers. While self-described as appropriate for “group care programs,” the SACERS has been used in a range of program environments beyond child care centers, including school-based after-school programs and community-based organizations such as YMCAs and Boys and Girls Clubs.

The SACERS, published in 1996 but updated periodically since then, is one of a series of program assessment instruments developed by researchers affiliated with the Frank Porter Graham Child Development Institute (FPG). As such, the SACERS is an adaptation of the Early Childhood Environment Rating Scale (ECERS) and is quite similar in format and mechanics to the ECERS, the Family Day Care Rating Scale (FDCRS) and the Infant/Toddler Environment Rating Scale (ITERS). Some states and localities have used several scales within the series to create continuity across accreditation or accountability systems, given the consistent orientation, language, format and scoring techniques.

Content

The SACERS measures process quality as well as corresponding structural features of programs. Its content reflects the notion that quality programs address three “basic needs” of children: protection of their health and safety, positive relationships and opportunities for stimulation and learning. These three basic components of quality care are considered equally important. They manifest themselves in tangible,

observable ways and constitute the key aspects of process quality included in the SACERS. The seven sub-scales of the SACERS include:

- Space and Furnishings;
- Health and Safety;
- Activities;
- Interactions;
- Program Structure;
- Staff Development; and
- Special Needs Supplement.

By addressing both process quality as well as structural features that relate to process quality (and other structural matters not directly related to process quality such as health policy), the SACERS puts as much emphasis, if not more, on program resources and the organization of those resources as it does on social processes that occur within the setting. This reflects its roots in the assessment and monitoring of environments serving young children. There are items on the SACERS that address each of the features of positive developmental settings outlined by the National Research Council (2002), with the most emphasis (the largest number of relevant items) clustering under the “physical and psychological safety” feature.

Interactions Sub-Scale Items

- Greeting/Departing
- Staff-child Interactions
- Staff-child Communication
- Staff Supervision of Children
- Discipline
- Peer Interactions
- Interactions Between Staff & Parents
- Staff Interaction
- Relationships Between Program Staff & Classroom Teachers

Structure and Methodology

The structure of the SACERS is straightforward and consistent with the other tools in the Environment Rating Scales series. The scale includes 49 items in the seven subscales mentioned above (see box for the items in the “Interactions” sub-scale). All of the sub-scales and items are organized into one booklet that includes the items, directions for use and scoring sheets.

While observation is the main form of data collection the instrument is built around, there are several items that are not likely to be observed during program visits. While the SACERS does not separate those items out into a separate interview scale or form, raters are encouraged to ask questions of a director or staff person in order to rate these items and are provided with specific sample questions that will help them get the necessary information to complete the form.

All 49 items are rated on a seven-point scale, with one being “inadequate” and seven being “excellent.” Concrete descriptions of what each item looks like

at a one, three, five and seven are provided (see examples below). Notes for clarification that help the user understand what they should be looking for and are also provided for many items. Observers compile their scores onto a summary score sheet, which encourages users to compile ratings and create an overall average program quality score.

The SACERS is meant to be used while observing one group at a time, for a period of three hours. A sample of one-third to one-half of groups (when programs have children divided into groups or classrooms) is required to establish a score for an entire program.

Technical Properties

In the case of the SACERS, psychometric evidence demonstrates that observations by different raters are consistent (interrater reliability) and that the instrument’s scales consist of items that cluster together in meaningful ways (internal consistency). Preliminary evidence also exists for concurrent validity, suggesting the SACERS may be an accurate measure of

	<i>Inadequate</i>		<i>Minimal</i>		<i>Good</i>		<i>Excellent</i>
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>
<i>Staff-Child Communications</i>	<ul style="list-style-type: none"> Staff-child communication is used primarily to control children’s behavior & manage routines. Children’s talk not encouraged. 		<ul style="list-style-type: none"> State initiate brief conversations (e.g. ask questions that can be answered with yes. no, limited turn-taking in conversations). Limited response by staff to child-initiated conversations & questions. 		<ul style="list-style-type: none"> Staff-child conversations are frequent. Turn-taking in conversation between staff & child is encouraged (e.g. staff listen as well as talk). Language is used primarily by staff to exchange information with children & for social interactions. Children are asked “why, how, what is” questions which require longer more complex answers. 		<ul style="list-style-type: none"> Staff make effort to talk with each child (e.g. listen to child’s description of school day, including problems & successes). Staff verbally expand on ideas presented by children (e.g. add information, ask questions to encourage children to explore ideas).

program practices that predict related outcomes.¹⁵ The information presented here is reported by Harms, Jacobs and White (1996).

Interrater Reliability

To examine interrater reliability or the degree to which different raters agree when observing the same program, paired raters assessed 24 programs using the measure. Researchers tested interrater reliability with the SACERS scales and total score using kappa scores and intraclass correlation coefficients.

All reliability coefficients were near or above 0.70, suggesting strong agreement. In other words, with adequate training for raters, scores will not depend on which rater is evaluating a given program.

Internal Consistency

Researchers examined how consistent individual item scores are within each respective SACERS scale, since all of the items within a particular scale are intended to measure a particular concept (e.g., Health and Safety). Internal consistency of the scales and the total score was strong, with alpha values ranging from .67 to .95. High internal consistency strengthens the argument that the items jointly represent the central concept of interest.

Convergent Validity

Convergent validity is examined by comparing the findings from the instrument of interest to a similar assessment tool, to help demonstrate the instrument's ability to measure what it is supposed to measure. Findings from three of the SACERS scales were compared to ratings with Vandell and Pierce's Program Quality Observation Scale (by the authors and colleagues of the PQQ, also reviewed in this report). Evidence indicated that each of these three SACERS scales (Program Structure, Activities and Interactions) were related to similar PQQ items in expected ways. Specifically, Vandell and Pierce (1998) found the following relationships between PQQ and SACERS

scales in 46 after-school programs: (1) SACERS Program Structure was positively related to PQQ Programming Flexibility, (2) SACERS Activities was positive related to PQQ Available Activities, and (3) SACERS Interactions was positively related to PQQ Staff Positive Regard and Positive Behavior Management, and negatively related to PQQ Staff Negative Regard and Negative Behavior Management. Convergent validity evidence is unavailable on the other SACERS scales.

Concurrent Validity

To determine whether SACERS accurately measures program quality, developers examined whether the instrument's ratings were related to distinct, theoretically important concepts in expected ways. Additional concurrent validity evidence covers all of the scales and total score. Specifically, because prior research suggests program quality is related to staff education/training, researchers expected that if the SACERS scales were adequately measuring quality, they would be positively related to staff education/training. As Harms, Jacobs and White (1996) expected, Space and Furnishings, Interactions, and Program Structure, as well as the overall SACERS score (which can be thought of as general program quality) were moderately, positively correlated with a measure of staff education and training. However, they did not report parallel correlations with measures of three additional scales (Health and Safety, Activities, or Staff Development); it is unclear whether they did not test these scales or if they found them to be unrelated to staff education/training. The researchers also tested the validity of the scales by examining their relationship to staff-child ratio. As expected, they found that Health and Safety, Activities, and Staff Development were moderately related with child-staff ratio. They did not report correlations between the other scales and total score with staff-child ratio and it is unclear whether they did not test these or whether they were uncorrelated with staff-child ratio.

Additional Validity Evidence

To explore the extent to which the SACERS adequately measures program quality, the developers asked nine

¹⁵ Except when noted, psychometric information is not available for the supplementary "special needs" items at the end of the instrument because none of the programs tested had exceptional children.

experts to rate how much each item in the instrument related to their definition of high quality. Using a five-point scale (with five being a very important aspect of quality), the minimum average score was around a four and experts rated most items close to a five. These scores suggest that items adequately measure aspects of quality. However, since experts were not asked whether any aspects of quality were absent from the instrument, this should not be taken as evidence that program quality as a whole is adequately represented.

User Considerations

Ease of Use

The SACERS is very easy to use in terms of accessibility of format and language (it is currently available in English, French, German, and most recently, Spanish). Full instructions for using the scale are included in the booklet along with the items themselves, notes clarifying many of the items and a training guide with advice on preparing to use the scale, conducting a practice observation and determining interrater reliability. One blank score sheet is included in the booklet and additional score-sheets can be ordered in packages of 30. The SACERS booklet is available for purchase through Teachers College Press at \$15.95.

Developers suggest it takes approximately three hours to observe a program and complete the form (users are encouraged to check off indicators and make at least initial scoring decisions while observing). Acknowledging that quality can vary within the same center or program, the developers advise that the approach to observation and scoring reflect how programs are structured. If a program has children broken into several different classrooms, observers are encouraged to observe one-third to one-half of the groups in the program before creating an overall score.

Available Supports

Three and five-day training workshops focused on the structure, rationale and scoring of the SACERS are available through the FPG Institute, as is additional information about the instrument and the other rating scales in the series. Specific guidance for how to conduct

your own training with staff or other observers is provided in the SACERS booklet. Training to reliability takes an estimated 4-5 days, with reliability checks throughout.

FPG is currently soliciting input from users in the field to develop a practical manual for adult educators using any of the rating scales, which will include specific materials such as course syllabi and outlines. Forms have also been developed to assist with reporting and applying observations to program improvement plans. Users can sign up to join a listserv through the FPG Web site to interact with other users in the field and to hear about updates and other relevant developments.

Large scale users of the rating scales can now work with a commercial software package – the ERS Data System – to enter and score their data. The Tablet PC version displays the items as seen in the print version and scores are made by tapping on the screen. Notes can also be written with a special pen and are automatically translated into print text and can be incorporated into the summary reports. The software also has a module on interrater reliability which can be used to compare scores, reach consensus and determine reliability. Using the Web-based system, individual assessments can be automatically routed to a supervisor for quality assurance and feedback and aggregate data analysis and organization and program-level reporting can be provided.

Important information for updating the SACERS is available at www.fpg.unc.edu/~ecers, including additional Notes for Clarification and an expanded scoresheet. Also, a revision of SACERS is forthcoming, as is a Youth rating scale for programs serving middle and high-school age youth.

In the Field

The state of Tennessee passed legislation in 2001 requiring all licensed child care centers and family/group homes in the state to be assessed using the Environment Rating Scales (including SACERS). The resulting Child Care Evaluation and Report Card program has two parts, one mandatory and one voluntary, both

of which are structured around the Environment Rating Scales to assess the quality of care provided at specific facilities. In the mandatory part of the program, the ERS assessment is one of several components of an overall “report card” given to each provider that must be posted along with their annual license.

The voluntary part of the program ties the ERS-based assessment to reimbursements. In the Star-Quality Child Care program, overall assessment scores for participating providers is converted into one, two, or three stars, which in turn can increase the provider’s state reimbursement by 5, 10 or 15 percent respectively. To support participation in both the mandatory and voluntary programs, local Technical Assistance Units provide assistance, at no charge, to any provider that wants information on how to improve quality and thereby increase its assessment score.

The Tennessee Department of Human Services (TDHS) works with the University of Tennessee and several other organizations to implement and manage this program. TDHS and UT’s Social Work Office of Research and Public Service manage the program and Tennessee State University prepares and delivers the initial training for assessors. Eleven resource centers around the state house an Assessment and Technical Assistance Unit. These units, which are responsible for conducting all the ERS-based assessments, hire and employ about 60 assessors statewide. Assessors receive ongoing training and frequent reliability checks by assessment specialists at the UT.

The assessment process takes place in conjunction with license renewal. A database has been developed that provides access to regularly updated statistical and demographic information about the program’s success in promoting, supporting and increasing quality child care across the state.

The SACERS and other scales in this series are part of many other state quality rating systems, including North Carolina, Mississippi, Arkansas and Pennsylvania.

For More Information

Additional information about the SACERS, supplementary materials and ordering information is available online at: www.fpg.unc.edu/~ecers/

Contact

Thelma Harms, Director of Curriculum Development
Frank Porter Graham Child Development Institute
517 S. Greensboro Street
Carborro, NC 27510
919.962.7358
harms@mail.fpg.unc.edu

Youth Program Quality Assessment

Developed by the David P. Weikart Center for Youth Program Quality¹⁶

Purpose and History

The Youth Program Quality Assessment (YPQA) is an instrument designed to evaluate the quality of youth-serving programs. While its practical uses include both program assessment and program improvement, its overall purpose 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 quality assessment tools and processes focus on the whole organization, 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. While some structural and organizational management issues are included in the instrument, it focuses primarily on those features of programs that can be observed and that staff have control over and can be empowered to change. While these social processes have not always been emphasized in licensing and regulatory processes, research suggests they are critical in influencing program quality and outcomes for youth. Given this focus, the YPQA is expected to assess program quality most accurately when users observe program offerings (programmatically consisting of the same staff, children and learning purpose across multiple sessions).

The YPQA has its roots in a long lineage of quality measurement rubrics developed by the High/Scope Educational Research Foundation over the past several decades for pre-school, elementary and now youth programs. In its initial iteration, the instrument was developed specifically to assess implementation of the High/Scope participatory learning approach. In its current form, the tool is relevant for a wide range of community- and school-based youth-serving settings that serve grades 4–12. It has been used in a range of after-school, camp, youth development, prevention and juvenile justice programs. It is not necessarily appropriate for use in highly unstructured settings that lack facilitated activities.

¹⁶ The Weikart Center is a joint venture between the High/Scope Educational Research Foundation and the Forum for Youth Investment.

Content

The YPQA measures factors at the Program Offering level and the Organizational level that affect quality at the “point of service.” The seven major domains (called sub-scales in the tool) that are covered include Engagement, Interaction, Supportive Environment, Safe Environment, Youth-centered Policies and Practices, High Expectations and Access.

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.

The content of the YPQA aligns well with the National Research Council’s features of positive developmental settings (2002), with the least emphasis on what is referred to by the NRC as “integration of family, school and community efforts.” The content of the YPQA has also been reviewed against and appears compatible with Jim Connell and Michelle Gambone’s youth development framework (2002).

Structure and Methodology

The seven topics or domains covered by the YPQA are measured by two different overall scales (groups of related items) that require different data collection methods. The program offering items are included in Form A and are assessed through observation. Form B includes the organization level items, which essentially assess the quality of organizational support for the program offering level items that are the focus of Form A. Evidence for Form B is gathered through a combination of guided interview and survey methods.

The seven domains can be graphically represented by the “pyramid of program quality,” (see below), which represents both an empirical reality and a unified

framework for understanding and improving quality. From an empirical perspective, assessments using the YPQA thus far follow a distinct pattern – most programs score highest in safety and then progressively lower as you move up the levels of the pyramid through support, interaction and engagement. Programs that score high in engagement and interaction appear most

able to influence positive youth outcomes (see technical properties for more detail on the validity study).

The scale used throughout the YPQA is intended to capture whether none of something (1), some of something (3) or all of something (5) exists. For each indicator, very concrete descriptors are provided to



illustrate what a score of 1, 3 or 5 looks like (see example on next page). The scoring for Forms A and B is consistent, but in the case of Form B, evidence to drive the scoring is based on an interview as opposed to observations. Observers are encouraged to write down evidence or examples that support the score that has been applied.

Technical Properties

Extensive psychometric evidence about the YPQA is primarily available from three studies. The first, referred to as the Validation Study, examined the reliability and validity of the instruments' scales with a sample of 59 organizations, most of which were after-school programs (Smith & Hohmann, 2005). The findings suggest the instrument has many good psychometric properties. Three of the seven scales, however, did not perform well in one or more psychometric areas.

The second study, referred to as the Self-Assessment Pilot Study, included a sample of 24 sites and specifically examined the YPQA's use as a self-assessment tool for after-school programs (Smith, 2005). This is the only study mentioned in this report that asked programs to assess themselves rather than

relying on independent researchers to collect data. This study examined the concurrent validity of the YPQA and found preliminary support for the total score and several scales. Similar to the first study, some scales exhibited problems with internal consistency.

The third study, referred to as the Palm Beach Quality Improvement System (QIS) Pilot Study, used a modified form of the YPQA known as the PBQ-PQA to assess program quality in 38 sites. The PBQ-PQA had similar, but not identical, scales compared to the YPQA (Smith, C, Akiva, T., Blazeovski, J. & Pelle, L., 2008).

In addition to these three studies, the developers also conducted additional interrater reliability analyses for the program offerings section of the instrument. They have also begun using techniques that provide more refined and detailed analyses of reliability and validity than traditional methods (see pages 16-17). In related work, CYPQ has just finished a validity study on a younger youth version of the PQA (grades K-4).

Score Distributions

Score distributions help users determine whether items adequately distinguish between programs on specific

II. Supportive Environment			
II-1. Staff Support Youth in Building New Skills			
Indicators			Supporting Evidence/ Anecdotes
1	3	5	
Youth are not encouraged to try out new skills or attempt higher levels of performance.	Some youth are encouraged to try out new skills or attempt higher levels of performance but others are not.	All youth are encouraged to try out new skills or attempt higher levels of performance.	n/o = 1
Some youth who try out new skills with imperfect results, errors or failure are informed of their errors (e.g., "That's wrong.") and/or are corrected, criticized, made fun of, or punished by staff without explanation.	Some youth who try out new skills receive support from staff who problem-solve with youth despite imperfect results, errors, or failure, and /or some youth are corrected with an explanation.	All youth who try out new skills receive support from staff despite imperfect results, errors, or failure; staff allow youth to learn from and correct mistakes and encourage youth to keep trying to improve their skills.	n/o = 1

dimensions. Smith and Hohmann (2005) examined average scores and spread for each of the scales and total scores for the Program Offerings and Organization items and found that all of the scales and total score had good distributions except for Safe Environment and Access (which each had means of 4.4 out of a possible 5.0). Most programs scored very high on these scales, making it hard to capture reliable differences. For Safe Environment, it may be realistic to assume that nearly all programs are relatively safe, particularly since the scores from this scale were validated by findings from a youth survey (see section on concurrent validity). However, additional evidence is needed to determine whether nearly all programs are high on Access, or whether there are meaningful differences that are not being picked up because the items are “too easy.” In the latter case, the items could be revised to better capture differences between programs.

Interrater Reliability

Recent analyses suggest that the current version of the tool paired with improved training techniques produces moderate to high levels of interrater reliability. For the Program Offering items, High/Scope researchers have captured four paired-rater data sets over the past two years for a total of 32 rater pairs using live and video methods for testing agreement. One of these data sets was produced independently by the Children’s Institute at the University of Rochester. All raters used the current version of the YPQA. Researchers found that across the rater pairs there was an average of 78 percent perfect agreement at the indicator level, which translates to an average maximum kappa coefficient of .66, close to the .70 benchmark for high interrater reliability. Similarly, the average item-level maximum kappa for the Program Offering items was also high at 0.72.

Findings suggest that the current version of the tool paired with rater training produces acceptable levels of interrater reliability for three of the four scales in the Program Offerings section. Specifically, the Safety, Support, and Engagement scales had acceptable reliabilities ranging between 0.66 and 0.73. The Interaction scale had moderate reliability (0.54).

Information for the Organization items (scales five through eight) comes from an earlier validation study by Smith and Hohmann (2005). The authors compared pairs of raters who examined the same programs at the same points in time. They examined the percentage of agreement across these items and found that the highest possible Kappa was 0.68, very close to the .70 benchmark for high reliability.

Smith and Hohmann (2005) also examined interrater reliabilities of the three Organization scales, which is important because users will ultimately draw most of their conclusions from the scales, not the individual items. They examined agreement using a statistic known as the intraclass correlation coefficient (ICC), which examines the degree to which differences among all ratings have to do with the difference between raters or differences among the programs themselves. The Youth Centered Policies and Practices, High Expectations and Access scales all had high interrater reliability (ICC = .51, .90 & .73 respectively).

Internal Consistency

Internal consistency indicates how closely related scores are for theoretically similar items. The Validation Study found that most of the YPQA scales exhibited acceptable internal consistency except for Safe Environment and Access. As noted above, this may have to do with the distributions of scores. Two items from an internally consistent scale go together, so that when item A is rated as high, item B is rated as high and when A is low, B is also low. However, if A is always high (because all programs do well on it), whether or not B is high, internal consistency will be low.

One example of an item in which most organizations received the highest possible score in the Self-Assessment Pilot Study was, “The physical environment is safe and healthy for youth.” If items such as this one are always high, we may not need to keep measuring them. However, if researchers believe that there is meaningful variation among programs, then these scales may need additional revision before we can be confident that their scores reliably measure the concepts that

they are supposed to measure. Similarly, Smith (2005) found in the Self-Assessment Pilot Study that these two scales had low internal consistency, but it also showed low internal consistency for two other scales: Youth Centered Policies and Practice and High Expectations for All Students and Staff. A possible explanation is that staff participating in the Self-Assessment Pilot Study were only given one day of training, whereas trained raters in the Validation Study may have been given more.

One additional explanation for why internal consistency may have been lower on some scales could be that the concepts forming these scales are formative rather than reflective. As explained in the section on Additional Technical Considerations (page 16-17), internal consistency tests are only appropriate when items are reflective, meaning that they all reflect the same underlying concept. Such items are closely related to one another and each represents a unique “attempt” to measure the concept of interest. However, internal consistency should not be used when items are formative, meaning that different components together make up or form a coherent set. For example, the Safe Environment scale may be more formative than reflective. A program that provides healthy food and drinks (as assessed by one item) may not necessarily have appropriate emergency procedures and supplies present (another item on the scale). However, even though these two items tap different underlying concepts (nutrition, safety in emergencies) and may not be closely related, their combination provides an important index of how a program promotes safety and health.

YPQA developers have begun examining whether some YPQA scales are formative versus reflective, and they are currently exploring whether certain items can be combined to form new, reflective scales.

Test-Retest Reliability

The Validation Study examined how much scores changed on multiple ratings over a period of three months. Correlations between assessments ranged from 0.81 to 0.98, indicating that ratings do not fluctuate widely over short periods of time. Long-term stability

was not assessed, so we cannot offer any evidence on whether the YPQA is sensitive enough to detect long-term change.

Validity of Scale Structure

Each of the scales in the YPQA is supposed to measure a separate concept. A factor analysis examines which items are similar to each other and which are different. Smith and Hohmann (2005) conducted a factor analysis at both observation periods and found preliminary evidence that the Program Offering items (scales two through four) grouped together in ways similar to the scales. Safe Environment was not included in the factor analysis and the authors acknowledge that the factor analysis did not support their expectations until they removed these items. Without the Safe Environment items, findings indicated that Supportive Environment and Opportunities for Interaction overlap and may not be entirely distinct. Validity support was strong for the Organization items (scales five through seven), which generally grouped together according to the theorized structure of the scales.

Convergent Validity

One way to examine whether an instrument actually measures aspects of program quality is to compare its scores to measures of identical or highly similar concepts. The Validation Study tested convergent validity by comparing all YPQA scales except Access and High Expectations to similar scales on a separate youth survey. For example, the Supportive Environment scale was compared to a Belonging scale on the youth survey. Correlational evidence indicates that the YPQA is moderately to strongly related to findings from the youth survey. The YPQA total scores for the observation and interview scales were also related to the youth survey total score. These results are encouraging for establishing validity.

In the Palm Beach Quality Improvement System Pilot Study, researchers examined the relationship between youth perceptions of program quality to modified versions of the four YPQA Form A domain scales. Scales on this form were similar to the original scales, but

not identical. Authors found that youth perceptions of quality were related to the Interaction scale, but were unrelated to Safe Environment, Supportive Environment, and Engagement scales. Although this evidence is mixed, this validity evidence may not apply to the current YPQA scales since the YPQA and the instrument used in this study are not completely identical.

Concurrent Validity

Concurrent validity is established when an instrument's items and scales are related to distinct but theoretically important concepts that are measured in the same time period. The Validation Study measured the validity of the total program quality score (created by averaging the various scale scores) by examining its relationship to expert ratings of the programs that were being evaluated. Specifically, experts rated programs based on youth centeredness and availability of resources. It is reasonable to expect that if the YPQA is indeed measuring program quality, then the total score would be related to these two expert-rated concepts. Using Pearson correlations as a measure of relatedness, Smith & Hohmann (2005) found strong evidence that the YPQA total score is related to expert ratings for these two domains, lending additional support that the instrument is indeed measuring program quality. They also tested the validity of the global program quality scores by comparing programs with trained staff to programs without trained staff. As expected, the programs with trained staff had higher global quality scores than those without, again lending support that the instrument can validly measure overall program quality.

The Validation Study also examined how well the instrument was associated with student experiences assessed by a separate youth survey (Smith & Hohmann, 2005). The following relationships were examined between YPQA and youth survey scales: (1) YPQA total score with the youth survey measure of overall program experiences, (2) YPQA Engaged Learning with measures of giving back to the community, youth growth, interest in the program, and challenging experiences, and (3) YPQA Interaction Opportunities with a measure of decision making in the program.

The authors found strong evidence for concurrent validity in that all of their hypothesized relationships were supported except for two (the Engagement scale was not related to youths' interest in the program or challenging experiences). However, this evidence is limited in that theoretically important relationships involving Form A's Safe Environment and Supportive Environment scales and the three Form B scales were not examined.

The Self-Assessment Pilot Study examined concurrent validity by correlating findings from the Supportive Environment and Engagement scales and the Program Offerings total score with a youth survey measure of staff support. Findings indicated a strong relationship between Supportive Environment and the youth survey. The Engagement scale was related in expected ways to a measure of program governance on the youth survey and the Program Offerings total score was related in expected ways to academic support and peer relations. None of these relationships were statistically significant, perhaps because the sample size was so small (12 programs). Thus, these relationships should be considered promising but not definitive.

The Palm Beach Quality Improvement System Pilot Study evaluated the concurrent validity of modified versions of the four Form A scales by examining their relationships with two scales from a youth survey: positive affect and challenging experiences. Higher positive affect scores were related to higher scores on YPQA Interaction Opportunities, but it was unrelated to the other three scales. Higher scores on challenging experiences was related to higher scores on YPQA Engaged Learning, but it was unrelated to the other three scales. In addition to the results being mixed, validity may not apply to the YPQA scales since they were not identical to the ones used in this study.

The concurrent validity evidence is promising but limited at this point. Additional support is needed for several of the instrument's scales.

Variations in Quality Across Different Contexts

Program quality may vary across different contexts such as different offerings, and how many sessions children and youth have had with one another. It is important to know if an instrument is sensitive to these types of differences, because if so, then users will need to conduct observations across a range of contexts. For example, if quality scores vary across different types of activities within a program, then users will need to observe a wide range of activities to obtain a complete picture of quality.

Developers have begun examining how the YPQA performs across three important program components: individual offerings, the content of these offerings, and how many sessions the children and staff have had together. Developers also examined variation across two combinations of these components. For example, does quality for some offerings stay constant throughout the year whereas quality for other offerings improves from the beginning to the end of the year? Are quality scores relatively similar in certain content areas regardless of which agencies are being observed whereas quality scores in other content areas vary across agencies? Currently, evidence is only available for the Interactions scale. Findings indicated that in addition to detecting quality differences across agencies, the Interactions scale was sensitive to differences across various types of offerings and content. However, measured quality did not vary by the number of sessions children and staff had together or across different combinations of programmatic components. These findings suggest that users of the YPQA should conduct observations across different types of offerings and content areas to obtain an accurate Interactions score.

In addition, even though agreement among raters was acceptable in other studies (as indicated in the section on interrater reliability), the developers found reliable differences among ratings given by different raters. This suggests that even good reliability among raters does not mean that raters should be ignored – a finding that probably extends to all the instruments in this compendium.

No evidence on how quality varies across contexts is currently available for the other YPQA scales, and it is also possible that the instrument is sensitive to other differences besides the ones already examined (e.g., time of day).

User Considerations

Ease of Use

The YPQA was developed with and for both practitioners and researchers; as a result the language is accessible and the format and scoring process is user-friendly. The administration manual and the introductions to Form A and B offer users a summary of the purpose and benefits of the tool, definitions of key terms used (e.g., scale, sub-scale, offering, item) and clear steps that walk users through the observation and scoring process. While training is recommended, the manuals themselves are self-explanatory. A “starter pack” that includes an administration manual, Form A and Form B can be ordered online for \$39.95.

Users of the YPQA are encouraged to conduct a running record of what occurs during a relatively extensive program observation as opposed to capturing several short snapshots of programming, because developers believe activities have a certain flow that is important to try to observe. This is particularly important if the goal is to come up with a reliable and valid score for an individual program as opposed to aggregating a large sample of observations for research purposes. Developers estimate that generating a score for a program, based on both Forms A and B, takes a minimum of approximately six hours for a single staff person. Roughly four of those hours are typically spent observing/interviewing within the program and another two hours writing up and scoring the instrument.

Available Supports

In addition to an online training, the Weikart Center offers YPQA training periodically around the country (which will soon be available online). The one-day workshop, YPQA Basics, introduces the observation and evidence gathering method, familiarizes participants with each item and indicator and prepares staff to

conduct the program self-assessment method of evidence gathering and scoring. The two-day YPQA Intermediate workshop covers all the material from the one-day and gives participants substantial practice scoring the tool using written scenarios and video, brings participants to acceptable levels of interrater reliability and prepares staff to conduct the external assessment method of evidence gathering and scoring. The three-day workshop covers all the material from these two trainings and includes a site visit (during which the participants score a youth program) and an analysis of the scoring efforts.

In the past year, the Weikart Center has developed a set of management-focused trainings to assist site managers in leading their programs through a data-driven quality improvement process.

The Weikart Center also offers 12 youth development trainings that are aligned with the content of the YPQA. Following a self-assessment or evaluation process, for example, program directors can assemble a tailored staff training experience based on specific areas within the YPQA where the assessment showed work was needed.

An electronic “scores reporter” is currently available from the Weikart Center (and is free to those who purchase the instrument). A more sophisticated Web-based data management system, is currently under development. This will allow individual programs or networks to join, go online to enter and analyze data and see their results at various levels of aggregation.

In the Field

The Rhode Island State 21st CCLC program has partnered with the Center for Youth Program Quality in a multi-year quality assessment process using a customized tool based on the research-validated PQA. The Rhode Island Program Quality Assessment (RIPQA), through a joint partnership of the Rhode Island After School Plus Alliance, the Providence After School Alliance and the Rhode Island Department of Education, is currently used by after-school programs across the city of Providence and throughout the state, including

all 21st CCLC funded programs. Participating programs conduct an annual self-assessment using the RIPQA. To support their efforts, a Weikart Center-trained Quality Advisor works with programs to jointly observe program offerings with site staff and then work one-on-one with agencies to develop quality improvement plans based on those observations.

As an additional component of this effort, the Weikart Center has also conducted a randomized field trial to test out their full training model. Based on 100 interviews with site supervisors, researchers have found that engaging providers in the observation and reflection process has been well-received across the board. The quality advisor and site-based technical support has been a very important part of the process, especially for those providers with limited capacity. Aggregated system-wide quality data are used to design and coordinate system-wide professional development offerings around the needs that get surfaced through assessment.

According to Elizabeth Devaney, Director of Quality Initiatives at the Providence After School Alliance, the quality improvement effort has “strengthened our position and ability to attract public and private resources to grow the system, and is an important strategy for sustainability going forward.”

For More Information

Information about the YPQA and ordering information is available online at:

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

Contact

Charles Smith, Director
David P. Weikart Center for Youth Program Quality
Centennial Plaza Building
Suite 601
124 Pearl Street
Ypsilanti, MI 48197
734.961.6900
csmith@cypq.org

References

- Arbreton, A., Goldsmith, J. & Sheldon, J. (2005). *Launching literacy in after-school programs: Early lessons from the CORAL initiative*. Philadelphia, PA: Public/Private Ventures.
- Arbreton, A., Sheldon, J., Bradshaw, M., & Goldsmith, J. with Jucovy, L. & Pepper, S. (2008). *Advancing achievement: Findings from an independent evaluation of a major after-school initiative*. Philadelphia, PA: Public/Private Ventures.
- Birmingham, J., Pechman, E., Russell, C., & Mielke, M. (2005). *Shared features of high-performing after-school programs: A follow-up to the TASC evaluation*. Washington, D.C.: Policy Studies Associates, Inc.
- Connell, J., & Gambone, M. (2002). *Youth development in community settings: A community action framework*. Philadelphia, PA: Youth Development Strategies Inc.
- Durlak, J. & Weissberg, R. (2007). *The impact of after-school programs that promote personal and social skills*. Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.
- Harms, T., Jacobs, E., & White, D. (1996). *School-age care environment scale*. New York, NY: Teachers College Press.
- Intercultural Center for Research in Education, & National Institute on Out-of-School Time (2005). *Pathways to success for youth: What works in afterschool: A report of the Massachusetts Afterschool Research Study (MARS)*. Boston, MA: United Way of Massachusetts Bay.
- Kim, J., Miller, T., Reisner, E. & Walking Eagle, K. (2005). *Evaluation of New Jersey After 3: First-year report on programs and participants*. Washington, DC: Policy Studies Associates, Inc.
- Knowlton, J., & Cryer, D. (1994). *Field test of the ASD program observation for reliability and validity*. Chapel Hill, NC: Authors.
- MacKenzie, S., Podsakoff, P., & Jarvis, C. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of Applied Psychology*. 90 (4). (pgs. 710-730).
- Martinez, A., & Raudenbush, S. W. (2008). Measuring and improving program quality: Reliability and statistical power. In M. Shinn & H. Yoshikawa (Eds.), *Toward positive youth development: Transforming schools and community programs* (pgs. 333-349). New York, NY: Oxford University Press, Inc.
- 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.
- Pechman, E., Mielke, M., Russell, C., White, R. & Cooc, N. (2008). *Out-of-School time observation instrument: Report of the validation study*. Washington, DC: Policy Studies Associates, Inc.
- Pierce, K. M., Bolt, D. M., & Vandell, D. L. (in press). Specific features of after-school program quality: Associations with children's functioning in middle childhood. *American Journal of Community Psychology*.
- Pierce, K., Hamm, J., Sisco, C., & Gmeinder, K. (1995). *A comparison of formal after-school program types*. Poster session presented at the biennial meeting of the Society for Research in Child Development, Indianapolis, IN.
- Pierce, K., Hamm, J., & Vandell, D. (1999). Experiences in after-school programs and children's adjustment in first-grade classrooms. *Child Development*. 70 (3), (pgs. 756-767).
- Raudenbush, S., Martinez, A., Bloom, H., Zhu, P., & Lin, F. (2008). *An eight-step paradigm for studying the reliability of group-level measures*. Chicago: University of Chicago.
- Russell, C., Mielke, M., & Reisner, E. (2008). *Evaluation of the New York City Department of Youth and*

Community Development: Out-of-school time programs for Youth Initiative: Results of efforts to increase program quality and scale in year 2. Washington, DC: Policy Studies Associates, Inc.

Russell, C., Reisner, E., Pearson, L., Afolabi, K., Miller, T., & Mielke, M. (2006). *Evaluation of the Out-of-School Time Initiative: Report on the first year.* Washington, DC: Policy Studies Associates.

Seidman, E., Tseng, V., & Weisner, T. (February 2006). *Social setting theory and measurement.* In William T. Grant Foundation Report and Resource Guide 2005-2006. New York, NY: William T. Grant Foundation.

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., Akiva, T., Blazevski, J., & Pelle, L. (2008, January). *Final Report on the Palm Beach Quality Improvement System Pilot: Model Implementation and Program Quality Improvement in 38 After-school Programs.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Smith, C., & Hohmann, C. (2005). *Youth program quality assessment youth validation study: Findings for instrument validation.* Ypsilanti, MI: High/Scope Educational Research Foundation.

Spielberger, J. & Lockaby, T. (2008). *Palm Beach County's Prime Time initiative: Improving the quality of after-school programs.* Chicago: Chapin Hall Center for Children at the University of Chicago.

Vandell, D. L., Reisner, E. R., & Pierce, K. M. (2007). *Outcomes linked to high-quality afterschool programs: Longitudinal findings from the study of promising afterschool programs.* Unpublished manuscript. Policy Studies Associates, Inc.

Vandell, D., Pierce, K., Brown, B., Lee, D., Bolt, D., Dadisman, K., Pechman, E., & Reisner, E. (2006).

Developmental outcomes associated with the after-school contexts of low-income children and youth. Unpublished manuscript.

Vandell, D., & Pierce, K. (2006). *Study of after-school care: Program quality observation.* Retrieved online at www.wcer.wisc.edu/childcare/pdf/asc/program_quality_observation_manual.pdf.

Vandell, D., & Pierce, K. (2001, April). Experiences in after-school programs and child well-being. In J.L. Mahoney (Chair). *Protective aspects of after-school activities: Processes and mechanisms.* Paper symposium conducted at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

Vandell, D., Reisner, E., Pierce, K., Brown, B., Lee, D., Bolt, D., & Pechman, E. (2006). *The study of promising after-school programs: Examination of longer term outcomes after two years of program experiences.* Wisconsin Center for Education Research, University of Wisconsin-Madison.

Vandell, D. L., & Pierce, K. M. (2001, April). Experiences in after-school programs and child well-being. In J. L. Mahoney (Chair), *Protective aspects of after-school activities: Processes and mechanisms.* Paper symposium conducted at the biennial meeting of the Society for Research in Child Development, Minneapolis, MN.

Vandell, D. & Pierce, K. (1998). *Measures used in the study of after-school care: Psychometric properties and validity information.* Unpublished manual, University of Wisconsin-Madison.

Walking Eagle, K., Miller, T., Reisner, E., LeFleur, J., Mielke, M., Edwards, S., Farber, M. (2008). *Increasing opportunities for academic and social development in 2006-07: Evaluation of New Jersey After 3.* Washington, DC: Policy Studies Associates, Inc.

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.

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.

Validity¹⁷

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).

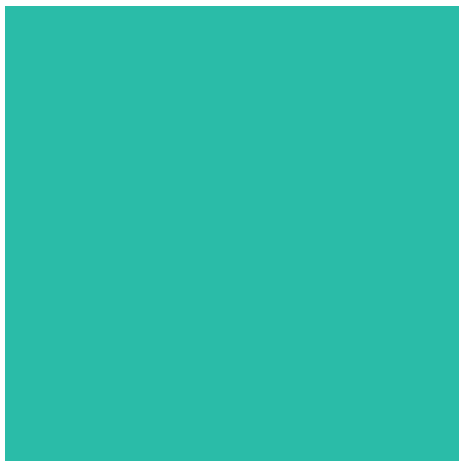
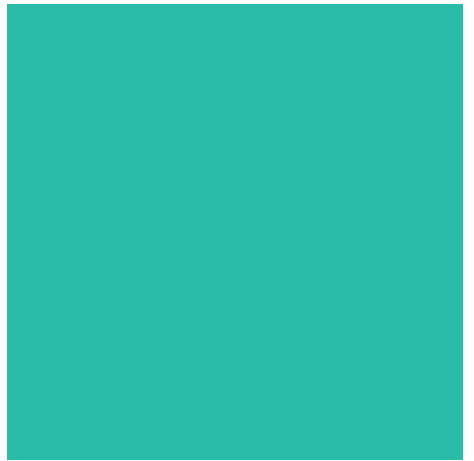
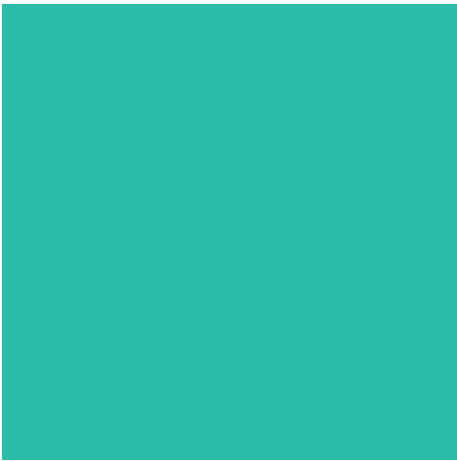
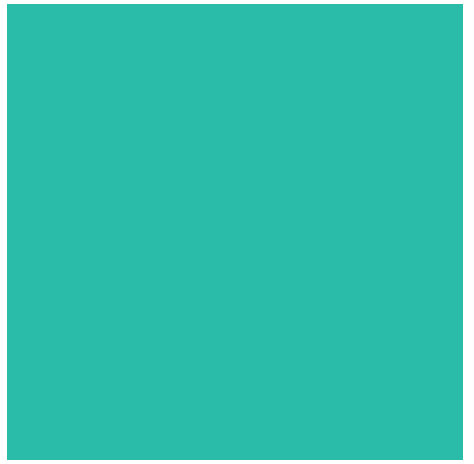
Factor analysis can also help determine if one scale actually incorporates more than one related concept. Imagine that we have an instrument with a scale called Homework Assistance, but our factor analysis finds that we actually have two separate concepts. We might discover that some items relate to 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

