The Assessment of
Afterschool Program Practices Tool (APT)

# Findings from the APT Validation Study

Allison Tracy, Ph.D.
Wendy Surr, M.A.
Amanda Richer, M.A.

NIOST | National Institute on Out-of-School Time
at the Wellesley Centers for Women

Wellesley Centers for Women

# Acknowledgements

The authors would like to thank the W.T. Grant Foundation for their generous funding and support for the *APT Validation Study*. Many individuals and organizations contributed to this study. We are especially grateful to our Research Associate, Ineke Ceder, for playing a crucial role in recruiting programs and helping to organize site visits and to our Research Assistant, Jillian Sherlock, for her many months of support to this study. We would particularly like to acknowledge and thank the administrators and staff from the 25 participating afterschool programs who dedicated their time and talent to the study and who graciously opened their doors to allow us to conduct observations at their sites, as well as to the youth who completed surveys as part of their site's participation. We would also like to thank the intermediary organizations that helped in the recruitment of sites for the study and in the provision of space for our training events: the Massachusetts Department of Elementary and Secondary Education, the United Way of Massachusetts Bay and Merrimack Valley, the YMCA of Greater Boston, the Lowell Public Schools, and the City of Cambridge Department of Human Services. Finally, we would like to extend our appreciation to the *APT Validation Study* observers, the local APT coaches/consultants, and NIOST researchers and administrators who participated in the Fall 2011 APT Forum to hear preliminary results from the study, offer suggestions, and provide feedback for the next steps in our work.

**Authors:**
Allison Tracy, Ph.D.
Wendy Surr, M.A.
Amanda Richer, M.A.

**Suggested Citation:**
Tracy, A., Surr, W., & Richer, A. (2012). *The Assessment of Afterschool Program Practices Tool (APT): Findings from the APT Validation Study*. Wellesley, MA: National Institute on Out-of-School Time.

# Executive Summary

The Assessment of Afterschool Program Practices Tool *(APT)*, developed by the National Institute of Out-of-School Time (NIOST), is an observational instrument designed to measure the aspects of afterschool program quality that research suggests contribute to the 21st century skills, attitudes, and behaviors youth need to be successful in school and the workplace. The APT is widely used across the country to support self-assessment and program improvement efforts. Increasingly, the APT is being used by external stakeholders for quality monitoring, assigning quality levels (e.g., as part of a Quality Rating and Improvement System), and identifying programs in need of improvement.

In 2010, with generous funding from the W.T. Grant Foundation, researchers from the Wellesley Centers for Women, Wellesley College, conducted the *APT Validation Study* with the aim of assessing the strength of the APT as a measurement tool. Based on observations of 25 afterschool programs serving grades K–8 in Massachusetts, this study provides scientific evidence that the APT possesses many strong technical properties. Among the study's many findings, researchers found that the APT captures key aspects of quality, such as whether a program is offering a welcoming environment or promoting youth engagement, which were found to be connected with positive youth program experiences and beliefs about themselves. APT ratings by a single observer are stable over time, which suggests that individuals can use the tool consistently to capture aspects of quality that are not overly sensitive to day-day fluctuations in practices. Practitioners do not tend to consistently rate their program higher or lower than outside observers who are unfamiliar with the program, suggesting that the APT can be used equally well by a variety of trained observers. Overall, the study suggests that the APT is an appropriate measure for examining afterschool program quality and is suitable for a number of lower-stakes purposes such as self-assessment and program support.

"The study suggests that the APT is an appropriate measure for examining afterschool program quality and is suitable for a number of lower-stakes purposes such as self-assessment and program support."

## *About the Assessment of the Afterschool Program Practices Tool (APT)*

The APT is one component of the Afterschool Program Assessment System (APAS), an integrated quality and outcome measurement system developed by the National Institute on Out-of-School Time (NIOST) in partnership with, and primarily funded by, the Massachusetts Department of Elementary and Secondary Education 21st Century Community Learning Center (MADESE-21st CCLC) initiative. The APT measures "process" quality – observable aspects of a program "in action" – through structured, live observations of a program across one or more afternoons.

APT development began in 2003 with a review of the research in the arts, education, and youth development literature to identify and define the program characteristics and staff practices that were associated with youths' acquisition of 21st century skills and behaviors. Subsequent testing of this earlier version of the APT showed preliminary evidence that the tool possessed desired technical properties (Yohalem & Wilson-Ahlstrom, 2009) and that the APT quality scores were associated with youth outcomes in expected ways (Miller, 2005). Between 2004 and 2005, the APT was adapted for use by afterschool programs for self-assessment and program improvement purposes. Extensive field testing of the APT by hundreds of MADESE 21st CCLC programs was conducted over a period of two years with the goal of creating a version of the tool that would be user-friendly and closely aligned with the needs of practitioners. The APT is currently used across the country to support self-assessment and program improvement efforts by both individual practitioners and by programs and intermediary organizations participating in city, state, and regional networks. Increasingly, the APT is being used by external stakeholders (e.g., funders and sponsors of afterschool programs) to monitor funded programs to ensure desired quality features, assign quality levels (e.g., as part of a Quality Rating and Improvement System), and identify programs in need of improvement.

*For more information on the APT please visit www.niost.org.*

# The APT Validation Study

This report describes the *APT Validation Study*, provides highlights of key findings, and explores implications and recommendations for using the APT to assess program quality. The *APT Validation Study* was designed to answer the following key research questions:

- ***Does the APT measure program quality in intended ways?***
- ***Does the APT produce consistent ratings of quality across raters and visits?***

Twenty-five afterschool programs serving youth in grades K-8 in the Greater Boston area participated in this study. Twenty-three afterschool staff practitioners and six external observers, unfamiliar with the programs, were sent to each of the afterschool programs twice, two weeks apart. Practitioners observed their own site for both visits, each time paired with a different external observer. Prior to site visits, all of the APT observers participated in an intensive two-day training and were instructed to follow specific research protocols. A diverse sample of 824 youth in grades 4-8 in these programs completed the Survey of Afterschool Youth Outcomes Youth Survey (SAYO-Y), responding to questions related to the quality of program experiences, their sense of competence both as a learner and socially, and their expectations and planning for the future. Researchers used these sources of data to perform a wide range of statistical tests to evaluate the APT as a measurement tool *(see the Appendix for more details on the design of this study)*.

**Summary of Evidence of the Technical Properties of the APT**

| Technical Property | What is it? | Strength of Evidence |
|---|---|---|
| Inter-rater Reliability | The extent to which raters agree in their ratings during the same observation segment. | ✔✔ |
| Test-Retest Stability | The ability of the tool to produce consistent ratings of a program's quality across multiple visits within a short time period. | ✔✔✔ |
| Score Range & Distribution | The range and dispersion of scores for a particular item or scale. | ✔✔ |
| Validity of Scale Structure | The degree to which individual items, when combined, can measure key areas of quality, broader domains of quality, and overall program quality. | ✔✔✔ |
| Concurrent/ Predictive Validity | The extent to which quality ratings are related to youth outcomes and produce a similar assessment of a program's quality as other instruments designed to measure comparable areas of quality. | ✔✔ |

Evidence of this property is...

✔ Weak or inconclusive *further research is needed*

✔✔ Mixed *further research is desirable*

✔✔✔ Strong *meets all/almost all benchmarks*

*Definitions for technical properties were adapted from the Forum for Youth Investment report of "Measuring Youth Program Quality: A Guide to Assessment Tools, Second Edition" (2009).*


# FINDING #1

## THE APT MEASURES KEY ASPECTS OF AFTERSCHOOL PROGRAM QUALITY IN INTENDED WAYS.

The APT is comprised of 72 items and 12 Quality Areas[1]. Each of the Quality Areas is measured by a group of items designed to work together to measure an aspect of quality that cannot be measured by a single item alone. As shown in Table 1 below, the APT Quality Areas fall into three broad Quality Domains: Supportive Social Environment, Opportunities for Engagement in Learning & Skill Building, and Program Organization & Structure.

| | Broad Quality Domains | | |
|---|---|---|---|
| | **Supportive Social Environment** | **Opportunities for Engagement in Learning & Skill Building** | **Program Organization & Structure** |
| **Quality Areas** | • Welcoming & Inclusive Environment<br>• Supportive Staff-Youth Relationships<br>• Positive Peer Relationships<br>• Relationships with Families | • Quality of Activities<br>• Staff Practices That Promote Engagement & Thinking<br>• Youth Engagement/ Participation<br>• Quality of Homework Support[2] | • Varied & Flexible Program Offerings<br>• High Program/ Activity Organization<br>• Positive Behavior Guidance<br>• Space Conducive to Learning |

*Table 1*

Of the 12 Quality Areas tested in this study, 10 met or exceeded accepted benchmarks for a strong scale structure.[i] In addition, the APT's Quality Areas were highly related to the three broad quality domains, as theorized. Furthermore, the study found that the APT ratings can be combined to produce an overall quality rating for the program. An unexpected finding was that the APT was also able to measure quality related to distinct times within the program day (e.g. arrival, transition, activity, homework, and pick-up time) by using the APT Time of Day sections (see Table 2 below). This finding is particularly important for practitioners and administrators who may opt to use the APT's Time of Day sections as a way to focus their self-assessment work.

1 One additional Quality Area – Program Promotes Autonomy and Leadership – was identified subsequent to this study and is not reported here. However, this scale underwent factor structure testing suggesting that it is strong and is associated with the Program Organization & Structure Quality Domain.

2 Quality of Homework Support is a new quality scale identified from the APT Validation Study. This scale includes all of the items (11 total items) from the Homework Time section of the APT.

<table>
<tr><th colspan="2">APT Time of Day Sections</th></tr>
<tr><td colspan="2">Observation Period</td></tr>
<tr><td>Informal Time</td><td>Structured Time</td></tr>
<tr><td>• Arrival<br>• Transitions<br>• Informal Program/ Social Time<br>• Pick-up</td><td rowspan="3">Activity Time<br>• Organization of Activity<br>• Nature of Activity<br>• Staff Promote Engagement & Stimulate Thinking<br>• Staff Positively Guide Youth Behavior<br>• Staff Build Relationships & Support Individual Youth<br>• Youth Participation<br>• Youth Relations with Adults<br>• Peer Relations<br><br>Homework Time<br>• Homework Organization<br>• Youth Participation<br>• Staff Effectively Manage Homework Time<br>• Staff Provide Individualized Homework Support</td></tr>
<tr><td>Academic Skill Building[3]</td></tr>
<tr><td>• Staff Promote & Encourage Academic Skills<br>• Youth Build & Practice Academic Skills</td></tr>
<tr><td colspan="2">Post-Observation</td></tr>
<tr><td colspan="2">• Program Space Supports Goals of Programming<br>• Overall Schedule & Offerings<br>• Overall Social-Emotional Environment</td></tr>
</table>

*Table 2*

> "An unexpected finding was that the APT was also able to measure quality related to distinct times within the program day (e.g. arrival, transition, activity, homework, and pick-up time) by using the APT Time of Day sections."

3 These sections of the APT are customized to meet the individual needs of the program and have not undergone psychometric testing.

# FINDING #2

## THE APT DISTINGUISHES BETWEEN THE QUALITY OF PROGRAMS AND THE QUALITY OF ACTIVITIES WITHIN A PROGRAM.

Often, funders, coaches, and technical advisors are working with multi-site organizations or program networks and would like to use assessment tools to compare quality across programs and identify programs in greatest need for support. The extent to which an observation tool is able to differentiate between the quality of sites is an important characteristic of a tool that is used for quality assessment. A series of statistical tests were performed to determine if overall site quality, Quality Areas, and Time of Day sections varied across sites in this study. Results show that the overall program rating and Quality Area ratings can be used to assess the quality of a given program relative to other programs; however, not all Time of Day ratings distinguish between sites.

Programs using the APT for self-assessment purposes are often interested in differentiating between the quality of activities offered at their site. Results showed that Activity Time ratings varied between activities within a site for most Quality Areas.[ii] This suggests that sites may be able to use ratings of specific activities to target those in need of improvement.

Most of the APT items, Quality Areas, and Time of Day sections showed high average ratings in this sample. This distribution of ratings is a characteristic shared with other observational instruments measuring quality of youth settings (Pianta, Paro, & Hamre, 2008; Smith & Hohmann, 2005). Given that the *APT Validation Study* included only a relatively small sample of programs, it is not clear whether the high average ratings were simply due to having a sample of higher quality programs or whether the APT tool contains items that represent quality practices that are easy for many programs to meet. Further study will be needed to answer this question.

> "Results show that APT overall program rating and APT Quality Area ratings can be used to assess the quality of a given program relative to other programs."

# FINDING #3

## APT RATINGS BY A SINGLE OBSERVER ARE STABLE ACROSS VISITS.

An important property of a quality assessment instrument is its ability to produce a stable rating of quality, capturing aspects of quality that are not overly sensitive to day-to-day fluctuations in practices. Results show that a practitioner observer's ratings of his/her own program are stable over the short term.[iii] This suggests that individual observers are able to use the tool consistently, and to capture stable aspects of quality across multiple days the program is visited.

***Important Note:*** The *APT Validation Study* did not find high agreement between independent external observers who assigned ratings on different days at the same program. Therefore, it is likely that external observers — such as coaches or assessors — may not be interchangeable across visits. Therefore, it is recommended that a consistent coach/assessor be assigned to a site.[iv]

# FINDING #4

## PROPERLY TRAINED PRACTITIONER AND EXTERNAL OBSERVERS CAN BE EQUALLY PROFICIENT RATERS. AGREEMENT ON THE RATING OF INDIVIDUAL APT ITEMS IS HARDER TO OBTAIN THAN SIMILARITY IN RATINGS FOR APT QUALITY AREAS AND TIME OF DAY SECTIONS.

One of the most important features of a quality assessment tool is its ability to produce accurate and objective ratings of quality that are free from variations due to subjective opinions, perceptions, and beliefs. In the *APT Validation Study*, observer pairs remained together during site visits and were required to follow a strict observation protocol prohibiting the sharing of notes, ratings, or discussion of impressions. Individual observers assigned their own independent ratings for each segment of the program day. Results show that most Quality Areas and Time of Day sections met minimum benchmarks for rater agreement but some did not meet more stringent benchmarks.[v]

Findings related to rater agreement for the individual APT items are mixed. The exact agreement was 59% on average across the APT items, and ranged from 21% to 100%. Few items passed statistical tests of rater agreement. Challenges related to reaching strong agreement between independent raters using similar observational instruments have been reported by other researchers (Bell et al., under review; Hill et al., 2012).

To explore the extent to which differences in ratings might be due to the characteristics of raters (e.g., age, gender, experience, education) or activity/observation conditions (e.g., length or type of activity, number of staff and youth present), a set of exploratory statistical tests were run.[vi] Results suggest few systematic rater differences due to rater characteristics and most activity/observation conditions. While raters are consistent overall, it appears that discrepancies in ratings may be due to individual interpretation of the APT items and personal differences in how raters applied the rating scale. To address these challenges, enhancements to rater preparation, training, and certification are needed.

> "Results show that most Quality Areas and Time of Day sections met minimum benchmarks for rater agreement but some did not meet more stringent benchmarks."

# FINDING #5

**PROGRAM QUALITY, AS MEASURED BY THE APT, IS RELATED TO YOUTH REPORTS OF SOME PROGRAM EXPERIENCES. MORE STRIKINGLY, THE APT QUALITY IS STRONGLY AND PERVASIVELY RELATED TO YOUTHS' ATTITUDES AND BELIEFS ABOUT THEMSELVES.**

Those interested in assessing program quality want to be confident that the aspects of program quality being measured are important to youth experiences and outcomes. Results show many associations between youths' program experiences and the APT ratings by Quality Area and by Time of Day.[vii] For instance, youth perceptions of having a supportive adult showed numerous connections with the APT ratings. Associations between the APT ratings and youths' attitudes and beliefs were even more prevalent and strong, with the most notable pattern being between the various APT ratings and youths' sense of competence as a learner. This combination of results suggests that the APT measures aspects of a program that are directly applicable to youth outcomes. Since findings were based on a small sample, further study is needed to confirm the association between APT ratings and youth outcomes.

> "Associations between the APT ratings and youths' attitudes and beliefs were even more prevalent and strong, with the most notable pattern being between the various APT ratings and youths' sense of competence as a learner."

# Implications for APT Use

The results of the *APT Validation Study* suggest that the APT can be used for a variety of purposes, by a variety of users, and in a variety of ways. The following table outlines the implications of study findings for these various assessment purposes. ***Important note:*** The *APT Validation Study* findings and implications for use assume that observers have undergone an intensive reliability training and have followed specific observation protocols.

| **Using APT for Self-Assessment & Program Improvement Support** | |
|---|---|
| Implications for Use | Limitations/ Recommendations |
| • Programs can opt to use the APT by Activity Offering, Time of Day, Quality Area, or broad Quality Domain to examine quality and identify areas/activities in need of strengthening.<br>• The APT ratings can be combined to create an Overall Quality rating for a site. Multi-site and intermediary organizations can use this rating to identify sites in need of support.<br>• The APT ratings can be used to help programs prepare for monitoring or other external assessment visits.<br>• A single, trained observer can gather consistent ratings for a program across multiple days.<br>• Trained observer pairs or teams visiting a program on the same day can split up and observe multiple activities and staff to produce a more comprehensive picture of program quality. | • It is <u>not</u> recommended that the ratings for a single APT item be used for making program improvement decisions. Programs should conduct multiple observations and combine ratings for items before selecting specific practices to strengthen.<br>• Observations of multiple activities are needed to help ensure a more accurate picture of a program's quality for self-assessment and program support purposes.<br>• The APT's ability to measure true change over time has not been evaluated.<br>• External observers may not be interchangeable from one visit to another. A consistent coach or Technical Advisor should be assigned to each site. |
| **Using APT for Monitoring & Assessment of Quality** | |
| Implications for Use | Limitations/ Recommendations |
| • Ratings of Quality Areas, broad Quality Domains, and/or Overall Site Quality scores can be used for reporting trends in quality.<br>• Quality scores can help stakeholders identify programs in need of support and strengthening.<br>• Quality scores can be used as part of monitoring to help determine the extent to which programs meet desired quality benchmarks as part of a low- or moderate-stakes effort.<br>• A consistent, trained assessor can gather quality data from a program across multiple days. | • APT ratings from a single site visit by a single observer <u>should not</u> be used for higher-stakes purposes.<br>• Multiple visits and ratings of multiple activities are highly recommended for a more accurate picture of program quality when the APT is being used for monitoring and other moderate-stakes purposes.<br>• External assessors may not be interchangeable from one visit to another. A consistent assessor should be assigned to each site. |

# Implications for APT Use

**Using the APT for Self-assessment**

The *APT Validation Study* findings support the strength and validity of the APT tool for internal, self-assessment purposes. The APT can provide practitioners with a new lens for looking at their program and to help identify strengths as well as areas to improve. Qualitative analysis of comments made by practitioner observers suggests that the APT is comprehensive, appropriate, and easy to use, and becomes even easier to use by a second site visit.

The study shows that the APT captures broad Quality Domains (Supportive Social Environment, Program Organization and Structure, and Opportunities for Engagement in Learning and Skill Building) and specific Quality Areas, as well as quality associated with program Times of Day (e.g., quality of arrival time, quality of homework time, etc.). APT ratings distinguish between the quality of activities within a program. This means that practitioners can use APT ratings to target specific activities and staff in need of improvement. Results of analysis show that the APT ratings can be combined to produce an overall quality rating for the program. Multi-site organizations can use overall site quality ratings to identify programs in need of support and strengthening.

Programs using the APT for self-assessment purposes have the option of examining quality by:

- Overall Program Quality
- Broad Quality Domains
- Specific Quality Areas
- Time of Day sections
- Differing activity offerings

The study shows that ratings by the same practitioner observer are consistent over time. This suggests that individual observers are able to use the tool consistently to capture stable aspects of quality across multiple visits. However, caution should be exercised before relying on ratings from a single visit to gain a picture of a program's quality. It is important to note that since the APT quality ratings differed across activities, those wishing to assess a program's quality at a broader level should be sure to observe and rate multiple activities during site visits. Additionally, the *APT Validation Study* did not examine whether the tool can be used over longer time periods to capture actual improvements (or declines) in quality. Further research is needed to confirm that the APT ratings can be used to accurately measure program improvement over time.

> "Since the APT is able to measure a range of key Quality Areas and can distinguish between the quality of activities within a program, practitioners can use the APT ratings to identify Quality Areas, distinct Times of Day, and specific activities and staff in need of improvement."

The study shows no evidence that practitioner raters tend to rate their programs consistently higher or lower than external raters. This suggests that practitioner observers, with proper training, can learn to use the tool and rate their program as well as someone from outside the program. Pairs or teams of observers, including both those familiar and unfamiliar with the program, can therefore observe differing program components to allow for a more comprehensive sample of program quality across multiple activities and staff.

Since the APT is able to measure a range of key Quality Areas and can distinguish between the quality of activities within a program, practitioners can use APT ratings to identify Quality Areas, distinct Times of Day, and specific activities and staff in need of improvement. However, since rater agreement was not high for the individual APT items, we recommend that average scores for Quality Areas and/or Time of Day sections be used, rather than individual item ratings alone, to make important decisions about program or quality improvement priorities. When programs wish to target specific practices contained in a single item, it is recommended that item ratings for multiple activities and from multiple observers be combined and discussed before selecting specific priority areas for improvement.

**Using the APT for Program Improvement Support**

Increasingly, the APT is being used by intermediary organizations, coaches, technical advisors, and funders to help identify and work with programs in need of support. Overall program ratings, as well as ratings for Quality Areas and some Time of Day sections can distinguish between sites. This allows funders, coaches and technical advisors working with multi-site organizations or program networks to use the APT to compare quality across programs and identify programs in greatest need for support. Coaches can encourage practitioners to use the tool themselves to self-assess their program, partner with practitioners to observe differing program components across multiple activities and staff to construct a more detailed and/or a more comprehensive assessment of program quality, and use the APT ratings to help programs prepare for monitoring or other external assessment visits.

The *APT Validation Study* findings support the strength and validity of the tool for use by those unfamiliar with the program such as a coach or technical advisor. However, high agreement was not found between independent observers who assigned ratings on different days. Therefore, it is likely that external observers such as coaches may not be interchangeable, particularly across time points. It is recommended that a consistent coach be assigned to a site when APT observations are used to determine areas to support. Coaches working with practitioner raters should observe multiple activities during each program visit to gain a more robust measure of program quality.

**Using the APT to Monitor and Assess Quality**

Increasingly, the APT is being used by external stakeholders such as funders and sponsors of afterschool programs to monitor funded programs to ensure desired quality features and to assess and assign quality levels (e.g., as part of a Quality Rating and Improvement System). The *APT Validation Study* was based on a small sample of programs, limiting our assurance that the APT, when used alone, can produce accurate enough quality ratings to be used for higher-stakes purposes. Future APT testing will be designed to assess the extent to which the APT and the training of the APT observers can help stakeholders produce highly accurate and consistent ratings of site quality necessary for higher-stakes purposes.

Study findings did show that the APT is suitable for quality assessment activities such as monitoring where ratings are used to identify program in need of further support. Since the APT Validation Study did not find high agreement between independent observers who assigned ratings on different days, it is suggested that external stakeholders use a consistent external assessor when observing a program across time points.

When the APT is being used for these moderate-stakes purposes or when information on program quality is being shared in any way, we suggest that the APT ratings be reported for Quality Areas or Time of Day sections rather than for individual items since the APT study findings did not show high rates of agreement at this level. In addition, it is recommended that raters reporting to external stakeholders observe multiple activities during each program visit and visit a program across multiple days to gain a more robust measure of program quality. Finally, it is suggested that external stakeholders do not base higher-stakes decisions (e.g., Quality Rating and Improvement System levels, funding decisions) on a single APT score derived from a single observer on a single day but should triangulate results across observers, days, activities, and across multiple assessment instruments.

"The APT Validation Study was based on a small sample of programs, limiting our assurance that the APT, when used alone, can produce accurate enough quality ratings to be used for higher-stakes purposes."

# Conclusion

The *APT Validation Study* provides scientific evidence to suggest that the APT possesses many strong technical properties, is an appropriate instrument for measuring after-school program quality, and is suitable for a number of lower-stakes purposes such as self-assessment and program support. Study findings also suggest some limitations in the APT use, particularly for higher-stakes purposes. Furthermore, since the study was based on a small sample of programs, it will be important for any future research to explore the extent to which the APT can be used for the full range of program contexts and purposes for which the tool is likely to be used.

**References**


Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (under review).
An argument approach to observation protocol validity.

Hill, H., Charalambous, C., McGinn, D., Blazar, D., Beisiegel, M., Humez, A., Kraft, M., Litke, E. & Lynch, K. (2012, February).
The Sensitivity of Validity Arguments for Observational Instruments: Evidence from the Mathematical Quality of Instruction Instrument. Unpublished manuscript, Harvard University.

Miller, B. M. (2005). Pathways to success for youth: What counts in afterschool.
Arlington, MA: Massachusetts AfterSchool Research Study (MARS).

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). Classroom assessment scoring system (CLASS) manual, K-3.
Baltimore, MD: Paul H. Brookes Publishing Co.

Smith, C., & Hohmann, C. (2005). High/Scope Youth PQA technical report: Full findings from the Youth PQA validation study.
Ypsilanti, MI: High Scope Educational Research Foundation.

Yohalem, N. & Wilson-Ahlstrom, A. (with Fischer, S. & Shinn, M.) (2009). Measuring youth program quality: A guide to assessment tools (2nd ed.). Washington, D.C.: The Forum for Youth Investment.

# Appendix: APT Validation Study
# Key Research Questions and Study Design

**Does the APT measure program quality in intended ways?**

- Is the APT able to measure specific program areas of quality? Broader domains of quality? Overall site quality? In what other ways can the APT be used to examine program quality?
- Are APT ratings able to distinguish between sites of varying levels of quality? Between the quality of differing activities offered within a site?
- Do APT ratings align with youths' own ratings of the quality of their program experiences?
- Are APT ratings of quality associated with positive outcomes for participating youth?

**Does the APT produce consistent ratings of quality across raters and visits?**

- To what extent do two independent APT observers visiting the same program assign similar ratings of program quality?
- Do afterschool practitioners observing their own program rate their program more leniently or severely than external observers who are unfamiliar with the program?
- To what extent do APT ratings assigned by a single observer remain consistent across visits taking place on differing days?

## Study Sample

**Sites:** Twenty-five afterschool programs serving youth in grades K-8 Greater Boston, Massachusetts area were recruited for the study. Participating sites represented a variety of program models, including 12 school-based programs, 4 community-based non-profits, and 9 sites affiliated with national organizations such as the YMCA and Boys and Girls Clubs of America. Almost half of the participating sites received 21st CCLC funding. Participating programs served varying age groups including those serving elementary only, middle school only, and K-8.

**Youth:** A diverse sample of 824 youth participated in the study by completing an online survey. Nearly equal numbers of males and females completed the survey and youth ranged from grades 4-8 with slightly more than half (65%) of the youth representing grades 4-5.

## Data Collection

**Site Visits:** For the study, pairs of observers (one practitioner and one external) were sent to each of the 25 afterschool programs twice, two weeks apart. Practitioner observers visited their own site for both visits, each time paired with a different external observer. External observers visited a different afterschool program site each time they observed. During site visits, observer pairs remained together, following a strict observation protocol, and assigned separate ratings for each segment of the program day including arrival, transition, homework, pick-up time, and at least two separate activity offerings. On the second visit, the same protocol and site visit schedule was followed, including observations of the same activity offerings observed during the first visit.

**Youth Responses:** At each site, all youth in grades 4-8 were invited to complete the Survey of Afterschool Youth Outcomes Youth Survey (SAYO-Y) within three weeks of the first APT observation visit. This online survey addresses the quality of program experiences, sense of competence both as a learner and socially, and expectations and planning for the future.

## APT Observers

Two types of APT observers were recruited for the study. First, an observer was selected from each participating site or organization. These 23 practitioner observers represented a variety of afterschool professional roles and had differing amounts of experience, background, and familiarity with the operations and youth at their site – ranging from direct care practitioners to Site Coordinators, Program Directors, and others. Next, six external observers were recruited. These individuals all had a background in the field of education and/or afterschool but were not familiar with the afterschool programs they would be observing.

## APT Observer Training

All of the participating APT observers were required to participate in an intensive two-day APT training prior to observing sites. The APT training sessions were highly interactive and utilized multiple methods for building observer's knowledge, skills, and ability to use the APT tool appropriately, as well as follow research protocols. Training included specific exercises to help minimize observer bias and games and other exercises designed to increase familiarity and facility with using the tool. Training also included DVD clips of actual programs in action, as well as participation in a live 90-minute, practice field visit at a nearby afterschool program.

# End notes

i Successful scales had more than two indicators, all with a minimum factor loading of .50.

ii Multilevel structural equation models (ML SEM) were used to examine quality of activities within sites, as well as average quality across sites.

iii Three tests were used to assess stability for individual items: Kappa, multifaceted Rasch models (MFRM), and multi-level structural equation models (ML SEM). Similarly, three tests were used to assess stability for scales: intraclass correlation coefficient (ICC), MFRM, and ML SEM. As a rule, these tests provide strong evidence of test-retest stability for the APT. We set the following benchmark standards: Kappa >.40, ICC>.50, MFRM non-significant time facet variance, ML SEM non-significant within level variance (time nested within practitioner observers).

iv Results showed that ICCs for externals across two time points only met minimum benchmarks for fewer than half of the quality areas (4 out of 11) and fewer than half of the Time of Day sections (6 out of 15). Average ICC across all scales was .26 and the media was .36.

v Similar to the tests of stability, there were multiple tests of inter-rater reliability. In this case, however, all tests except MFRM were conducted at both Time 1 and Time 2. Conclusions are based on the weighted average of the test values across the two times. In addition to these tests, exact agreement was calculated, including ratings of N/A given for conditions not observed.

vi ML SEM regression analysis was used to estimate the degree to which rating disagreements was related to rater or activity characteristics.

vii Pearson correlation coefficients were calculated using the APT scale scores averaged across raters and occasions and SAYO responses averaged across youth in a program.