



Measuring Program Quality:

Evidence of the Scientific Validity of the Assessment of Program Practices Tool

Allison Tracy, Linda Charmaraman, Ineke Ceder, Amanda Richer, and Wendy Surr

Out-of-school time (OST) youth programs are inherently difficult to assess. They are often very dynamic: Many youth interact with one another and with staff members in various physical environments. Despite the challenge, measuring quality is critical to help program directors and policy makers identify where to improve and how to support those improvements.

This article describes recent research on the Assessment of Program Practices Tool (APT), establishing its strength as an evaluation and tracking tool for OST programs. Funded by the William T. Grant Foundation and Virginia B. Toulmin Foundation, the validation was conducted in two phases. The first phase was designed to evaluate the scientific rigor of the tool. Based on the findings from the first phase, the second aimed to inform improvements in the tool and its training. Our testing so far shows that online video-based training needs to be more specialized in order to improve rating reliability for high-stakes users, such as third-party evaluators.

ALLISON TRACY has over 15 years of experience providing methodological and statistical consultation for researchers in a wide variety of disciplines, topics, and institutions. She has worked on the psychometric analysis of A Program Assessment System (APAS) and its components, including the Assessment of Program Practices Tool (APT).

LINDA CHARMARAMAN, PhD, co-principal investigator of the APT Validation Study II, is a research scientist at the Wellesley Centers for Women specializing in positive youth development. She has conducted research and evaluations of in-school and out-of-school time (OST) programs for over 12 years.

INEKE CEDER is a research associate at the Wellesley Centers for Women, where she has been involved with projects on child and adolescent development, sexual education, and women's leadership. She was the data collection manager and survey developer for Phase 2 of the APT Validation Study.

AMANDA RICHER is a research associate for the National Institute on Out-of-School Time and assistant methodologist for the Wellesley Centers for Women. She has been involved in psychometric testing of OST assessments and she has supported research in youth development.

WENDY SURR, Senior Researcher at the American Institutes for Research, has more than 30 years of experience leading research and evaluation studies and other educational initiatives. While at NIOST, she co-created APAS, including developing the APT and a set of teacher, staff, and youth survey instruments for measuring student nonacademic outcomes. She served as co-principal investigator and project director for Phase 1 of the APT Validation Study.

The Assessment of Program Practices Tool

The APT is an observational instrument created in 2005 by Beth Miller and Wendy Surr of the National Institute on Out-of-School Time (NIOST) to measure OST *process quality*: observable aspects of a program in action. Research suggests that process quality contributes to the 21st century skills, attitudes, and behaviors youth need to be successful in school and the workplace (Miller, 2005).

The APT is one component of A Program Assessment System (APAS), an integrated quality and outcome assessment system developed by NIOST in partnership with the Massachusetts Department of Elementary and Secondary Education 21st Century Community Learning Centers (21st CCLC) initiative. Currently, the APT is used in 33 states and in Canada by over 1,500 individuals and 600 OST programs. The APT is designed to support program self-assessment and improvement efforts. Increasingly, it is also being used by external stakeholders, such as funders and sponsors of afterschool programs, to ensure that programs are implementing quality features. External observers are using the APT to assign quality levels, often as part of a quality rating and improvement system, in order to identify programs or program aspects in need of improvement.

The APT measures aspects of process quality in three key domains: supportive social environment, opportunities for engagement in learning, and program organization and structure. As shown in Table 1, these three domains have 12 subdomains called *quality areas*. The items measuring a given quality area might be drawn from different sections of the APT, which is laid out to follow specific program times: arrival, transition, homework, activity, informal, and pick-up.

Each item is rated on a 4-point scale, where 4 represents the desired practice. Detailed item-specific “anchors” define each rating point and provide observable indicators to guide scoring. Figure 1 illustrates the anchors for a *conditional* item: one that can be rated only if the condition, in this case youth misbehavior, is observed.

APT Validation Study Phase 1: Scientific Rigor

In 2010, Phase I of the APT Validation Study was conducted to assess the technical properties of the tool and confirm its scientific rigor. We conducted four tests:

- **Internal consistency:** whether all items assigned to the same scale receive a high score when a program shows high quality in a given quality area; whether the items as a set can distinguish higher- and lower-quality programs
- **Test-retest stability:** whether scores are not overly sensitive to day-to-day fluctuations in quality
- **Interrater reliability:** whether two raters observing the same program on the same day give the same ratings
- **Predictive validity:** whether programs with high APT scores have better youth outcomes than programs with lower APT scores

Methods

Participating Programs

We recruited 25 afterschool programs in greater Boston: 12 school-based programs, four community-based nonprofits, and nine sites affiliated with national organizations such as the YMCA and Boys & Girls Clubs of America. Almost half received 21st CCLC funding. Programs served varying age groups: elementary only, middle school only, and K–8. A diverse sample of 824 youth in grades 4–8, equally male and female, completed an online survey; slightly more than half (65 percent) were in grades 4–5.

Table 1. Three APT Quality Domains and Their Associated Quality Areas

Domain	Quality Areas
Supportive social environment	Welcoming and inclusive environment Supportive staff-youth relationships Positive peer relationships Relationships with families
Program organization & structure	Space conducive to learning Varied and flexible program offerings Positive behavior guidance High program activity organization
Opportunities for engagement in learning & skill building	Youth autonomy and leadership Youth engagement and participation Quality of homework support Staff practices that promote engagement & thinking

Figure 1. Sample APT Rating Anchors

When youth behavior is inappropriate, staff use simple reminders to redirect behavior.			
1	2	3	4
<p>Staff do not use simple reminders to redirect behavior OR always over-react to youth behavior.</p> <p>Example: Staff reactions to youth behavior are not instructive or constructive. Staff are often visibly frustrated, sarcastic, and short-fused with youth; e.g., saying loudly "How many times do I have to tell you to keep your hands to yourself?"</p>	<p>Staff sometimes use simple reminders to redirect behavior but over-react to youth behavior.</p> <p>Example: Staff are easily frustrated by youth and turn small incidents into a bigger deal than necessary, such as spending 10 minutes explaining why youth should keep their hands to themselves.</p>	<p>Staff usually use simple reminders to redirect behavior but over-react slightly to youth behavior.</p> <p>Example: Staff respond to too many mild youth behaviors such as reminding youth to keep their hands to themselves multiple times while they are waiting in line.</p>	<p>Staff always use simple reminders and are always calm when handling youth behavior.</p> <p>Example: Staff let youth know what is inappropriate and remind them of established rules and behavioral expectations.</p>

Participating Raters

Two types of APT observers were recruited. First, an *internal* observer was selected from each participating program. These 23 internal observers (one observer oversaw three of the study programs) were direct care practitioners, site coordinators, program directors, and others with varying backgrounds and levels of experience. Next, we recruited six *external* observers who all had backgrounds in education or afterschool but were not familiar with the afterschool programs they observed and had not been trained to use the APT.

All 29 observers participated in an intensive two-day training to learn to use the APT and to follow research protocols. Training included exercises to minimize observer bias, games and exercises to increase facility with the tool, extensive practice rating DVD clips of actual programs, and a 90-minute live practice field visit.

Data Collection

One external and one internal observer were sent to each of the 25 programs for two visits two weeks apart. Internal observers studied their own site for both visits, each time paired with a different external observer. External observers studied a different site each time. During site visits, observer pairs remained together but assigned ratings separately, following a strict observation protocol. At each site, all youth in grades 4–8 were invited to complete the Survey of Afterschool Youth Outcomes-Youth (SAYO-Y) within three weeks of the first visit. The SAYO-Y, a part of APAS, is a self-report instrument that measures three key areas: youths’ pro-

gram experiences, their sense of competence, and their future planning and expectations. Initially developed in 2008, SAYO-Y has undergone extensive testing to confirm its consistency and validity.

Analysis and Results

Analysis of the observation ratings and the corresponding SAYO-Y scores led to five findings about the consistency, stability, reliability, and validity of the APT.

Finding 1: APT items can be combined to create reliable scale scores.

The findings show that the APT items designed to measure the same quality area work together as a set to distinguish among programs with varying levels of quality and that items designed to measure one quality area are distinct from items designed to measure another quality area. Furthermore, the study found that APT items could be combined to produce an overall rating of quality.

We also established that the items in each APT section representing a specific time of day could be reliably combined into a scale score to assess the quality of, for example, homework time or activity time. This finding is particularly important for programs that opt to focus their self-assessment on particular times of the program day, rather than using the entire APT.

Finding 2: The APT can be used to compare programs or activities within a program.

A rigorous tool must be able to capture different levels of quality, from very low to very high. A tool is not useful if

it captures small distinctions among very high-quality programs but cannot distinguish between moderate- and low-quality programs. Therefore, we tested whether, among all the programs studied, scale scores ranged along the full continuum from very low to very high.

We used statistical models to examine variability in quality across sites and among activities within sites. Results show statistically significant differences in the quality of programs, confirming that the APT can be used to distinguish between one program and another either in overall quality or in specific quality areas. Scales created for specific time-of-day sections were less able to capture differences between or within programs—except for the activity time scale, whose scores captured statistically significant quality differences among activities within a site. Programs thus can use activity time scale scores for self-assessment.

Finding 3: APT scale scores are not overly influenced by program fluctuations.

A quality assessment instrument must produce stable quality ratings that are not overly sensitive to day-to-day fluctuations in practices. When a program is assessed during a short time window, real change in quality is not expected to occur, so the APT scores should be similar.

When we assessed test-retest stability for individual items, quality areas, and time-of-day scales, we found that internal observers produced ratings that were stable over the short term. Observers should therefore be able to use the APT to capture aspects of quality that are stable across multiple observation days.

Finding 4: Perfect interrater agreement is hard to achieve.

A quality assessment tool must be able to produce accurate quality ratings that are free from variations due to subjective opinions and perceptions. No matter who conducts the observation, a program's quality ratings should be the same.

When we tested interrater reliability for individual APT items, findings were mixed. The average rate at which both observers assigned the exact same rating was 59 percent; the range for all raters was 21 percent to 100 percent. Few items passed statistical tests of interrater agreement. Other researchers have reported similar challenges in reaching interrater agreement for similar observational instruments (Bell et al., 2012; Hill et al., 2012).

We used a set of exploratory statistical tests to explore the extent to which differences in ratings might be due to characteristics of raters, such as their age, gender, experience, and education, or to observation conditions,

such as the length or type of activity and the numbers of staff and youth present. We found that agreement was harder to attain in observations of sports and active games. This finding is not surprising: These activities can be fast moving and cover large spaces, so that observers could have trouble hearing and seeing interactions.

Though internal and external raters often disagreed, internal raters were consistent in their ratings of their own programs over time. Use of the APT as a self-assessment for program improvement therefore appears warranted. However, comparison of one program with another by external raters, especially when stakes are high, may require more training to produce better rating agreement.

Finding 5: The APT measures program aspects that are directly related to youth outcomes.

Those interested in assessing program quality want to be confident that the quality areas being measured are important to youth experiences and outcomes. To examine the concurrent and predictive validity of the APT, we analyzed the relationships between the quality areas and youth responses to the SAYO-Y. Results show many associations between APT ratings and youths' program experiences, as summarized in Table 2. For instance, youth perceptions of having a supportive adult show numerous connections with APT ratings. Associations between APT ratings and youths' attitudes and beliefs are even more prevalent and strong; the strongest correlation is between youths' sense of competence as learners and several APT quality areas.

Revisions

Based on the item-level results, some APT items were dropped or revised, and newly revised items and their anchors were piloted. These improvements were incorporated into the instrument for Phase 2 of the APT Validation Study.

APT Validation Study Phase 2: Training

Phase 1 findings suggest that, although individual raters are consistent over time, they do not always agree with other raters. Researchers have had the same result with similar observational instruments (Hoyt & Kerns, 1999; Lumley & McNamara, 1993).

Growing interest in use of the APT for high-stakes purposes, such as quality rating and improvement systems, led to interest in enhancing interrater reliability. Training and practice have been found to increase rater scoring accuracy (Hoyt & Kerns, 1999; Knoch, Read, &

Table 2. Relationships Between APT Quality Area Scores and SAYO-Y Scale Scores

APT Quality Areas	SAYO-Y PROGRAM EXPERIENCES					SAYO-Y ATTITUDES AND BELIEFS			
	Supportive Social Environment	Enjoyment & Engagement	Choice & Autonomy	Challenging Environment	Supportive Adult	Sense of Competence as a Learner	Sense of Competence Socially	Future Planning	Future Expectations
Welcoming environment						Light		Light	Light
Supportive staff-youth relations					Light	Dark		Light	Light
Positive peer relations	Light				Light	Dark	Light	Light	Dark
Space conducive to learning		Light			Light	Dark		Light	Light
Varied & flexible program			Dark			Dark	Light	Light	Light
Behavior guidance					Light	Dark	Light	Light	Light
High program activity organization			Light		Light	Dark	Light	Dark	Dark
Youth autonomy & leadership			Light			Light			
Youth engagement	Light	Light	Light		Light	Dark	Light	Light	Dark
Quality of homework support		Light		Light	Light				
Promotion of thinking & engagement				Light	Light				

Note: Dark shading signifies a *very strong* relationship (correlation of .50 or greater) and lighter shading signifies a *strong* relationship (correlations between .30 and .50).

von Randow, 2007; Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009). We therefore developed an enhanced APT training with three components:

1. The comprehensive APT Anchors Guide, which offers scoring criteria for each item (see Figure 1)
2. Master-scored online practice video clips with detailed rationales for the assigned scores
3. Advanced in-person training
4. Targeted feedback with recommendations for additional practice

Strict use of the APT Anchors Guide was intended to focus observers solely on the observable behavior of staff and youth. Enhanced training, including the video clips, was designed to minimize subjective interpretation and discrepancies among raters. Our primary research question was, “Do trainees who undergo APT training and practice improve in the accuracy of their APT ratings?”

Methods

Participating Trainees

We identified a sample of APT trainees to reflect the expected profile of likely APT raters with respect to geography, prior experience, age, and familiarity with APT. Our sample of 39 trainees was drawn from the New England area and from the South. The sample was 69 percent female and 59 percent White, 26 percent Black, and 15 percent non-Black minority. Twenty-six percent of trainees were under 30 years old, 36 percent between 30 and 40, 23 percent between 41 and 50, and 15 percent older than 50. The majority had experience with programs for elementary and middle school youth. Only 38 percent of trainees reported that they had ever used the APT Anchors Guide.

In order to evaluate the reliability training, we asked the trainees to complete four video-rating exams: one at baseline and one after each major component of the

training, as shown in Figure 2. Using ratings by master raters (“master scores”) as a reference, we examined the results to see whether trainees improved in the accuracy of their ratings and, if so, at what points in the training and for which APT sections or scales.

We chose to use video clips rather than live practice opportunities for training in order to ensure that the focus of the observation was consistent across raters. In live observations, the 360-degree view of the environment means that two observers may pay attention to different activities and therefore rate different sets of staff and youth. For training purposes, we needed to narrow the field of focus in order to draw trainees’ attention to specific instances that they could map onto the anchors for each item. Furthermore, using video clips enabled us to define accuracy as the match between trainee ratings and master scores, as opposed to the less precise method we used in Phase 1, where agreement between two raters served as a proxy for accuracy.

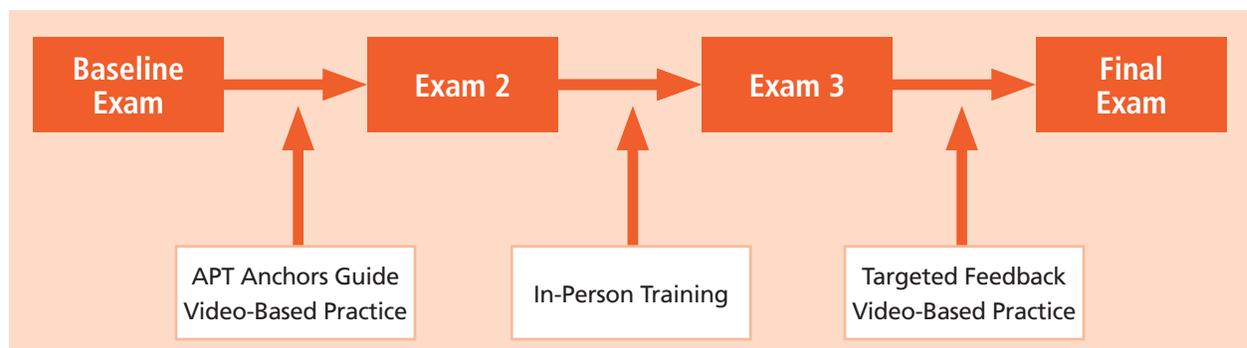
Training Components

The enhanced training, as outlined above, had three major components.

APT Anchors Guide. A key aspect of training was providing the comprehensive Anchors Guide in order to build raters’ familiarity with and use of the anchors.

Master-Scored Videos. To create video clips for use in the study we selected eight New England afterschool programs, based on considerations such as size, type, location, ages served, and race/ethnicity of youth served. To capture a variety of program practices, these programs were videotaped over four days. Each recording was subdivided into a number of shorter clips, organized by the time-of-day sections of the APT. The 350 resulting video clips ranged in length from one minute to 20 minutes. Each clip was reviewed by up to four master raters who

Figure 2. Trainee Exams and Training Components



had extensive experience in afterschool and were familiar with the APT. Clips were considered for inclusion in APT exams and practice modules if the audio and video quality was good and if agreement among master scorers was high. Furthermore, the clips represented various anchors and conditional items (for example, “if there is youth misbehavior”).

Each APT practice module and exam included one clip from each APT time-of-day section. All were approximately one hour long; included only one clip from any one of the eight recorded programs; and offered a good representation of low-, medium-, and high-scoring clips. Following each time-of-day video clip, the online program displayed the relevant APT items, with their lowest and highest scores, and trainees rated the video clip on each of the items in that APT section. Immediately afterward, the master score and the reasoning behind that rating were displayed. In practice modules, trainees could go back to review the video clips after seeing the master scores. The exams did not offer this option.

In-Person Training. After rating the videos, trainees participated in a six-hour in-person APT training event. Two highly experienced APT lead trainers focused the in-person training on the 15 APT items on which trainees had the lowest rates of agreement with master scores in the exams. The trainers used video clips from the exams to demonstrate common sources of ambiguity, such as interactions that fell between a rating of 2 and 3. Small- and whole-group discussions enabled the trainers to open a dialogue so trainees could come to a collective conclusion about the evidence supporting the master score. Agreeing about the evidence is a key step in improving accuracy.

Targeted Feedback and Additional Practice. After the in-person training and its exam, trainees were offered feedback recommending that they complete additional practice modules in the area in which they scored lowest.

Data Collection

The primary data collection instruments were video exams, one at baseline and one each after receipt of the APT Anchors Guide, after the in-person training, and after the targeted feedback. The final exam took place within three weeks of the in-person training. It included a qualitative process evaluation asking trainees about their experience with the training materials and their level of confidence in assigning APT ratings.

Analysis and Results

Between the baseline exam and exam 2, we asked trainees to rate at least one of the two practice clips for each APT time-of-day section, aiming for six practice clips. They rated an average of 4.6 clips. Between exams 3 and 4, participants rated an average of 2.15 clips out of the recommended four. Trainees reported varying levels of use of the APT Anchors Guide: 64 percent said they used it always, 31 percent some of the time, and 5 percent rarely. Participants who were White, female, older, or from the South completed more practice modules and referred more often to the guide.

Across all four exams, 53 percent of the trainees’ ratings matched the master scores exactly. Trainees were more likely to match the master scores when scores were at the high or low end of the rating scale and when the youth in the clip were in middle school rather than elementary school. White and non-Black minority trainees were more likely to match the master score than were Black trainees. In several time-of-day sections, trainees had more matches when the clips were shorter; only in the homework section did longer clips yield more matches. In the open-ended evaluations, a high percentage of trainees recommended using shorter video clips. The reasons they gave were attention span limits for videos, issues with narrow camera angles and audio quality, and a preference for focusing on a limited program snapshot.

In order to test the hypothesis that training would improve in the match between trainee ratings and master scores, we examined the results of exams 2–4. To do so, we made statistical adjustments to compensate for differences among the exams in such characteristics as the total quality score of all the clips, clip duration, and participant age group. We also adjusted for lack of compliance with the exam protocol, as when trainees cut the clip short rather than viewing to the end. This analysis resulted in four significant findings.

Finding 6: Video practice shows promise as effective training mode.

The process of rating practice videos using the APT Anchors Guide and receiving immediate feedback on the rationale for the master scores led to significant improvement in trainees’ rate of agreement with master scores, from an average of 49 percent on the baseline exam to an average of 52 percent on the second exam. One trainee noted in the qualitative process evaluation, “I have become more discrete in my ratings, and am much more comfortable using the anchors as a guide when providing evidence for why I rated a certain way.”

Finding 7: The in-person training did not improve average rating accuracy.

Exam 3, administered after the in-person training, showed a significant overall decline in trainees' rate of agreement with master scores, by an average of 6 percentage points. However, 16 of the 39 participants did not decline, and eight improved by 6 or more percentage points. Two participants experienced strong gains of 15 and 19 percentage points.

We sought but did not find characteristics that distinguished trainees who improved from those whose scores declined. However, participants' comments may shed some light. Many participants said that they found the in-person training helpful because they could ask questions and share insights with others. However, some participants did not find that the training improved their accuracy.

Finding 8: Targeted feedback and additional video practice did not further improve scores.

The final exam, administered after recommendations for targeted practice after exam 3, did not yield significant improvement in trainees' agreement with master scores. Analysis of changes in scores for time-of-day sections of the APT yielded similar results, with the greatest improvement emerging between the baseline and exam 2, after the video training. Trainees may have experienced a plateau effect, even though the highest rate of agreement with master scores on exam 3 was only 66 percent. Another explanation could be burn-out due to the heavy demands the study made on participants.

Finding 9: High-priority APT quality areas showed the most improvement.

Four of the APT program quality areas showed improvement in ratings: supportive staff-youth relations, positive peer relations, behavioral guidance, and high program activity organization. In all four areas, average trainee scores showed considerable improvement from baseline to exam 2, ranging from 8 percentage points for behavioral guidance to 20 percentage points for positive peer relations. Three of the four quality areas showed improvement in accuracy across the full training experience, from baseline to exam 4. The fourth area, supportive staff-youth relations, showed an average decline

in accuracy of 3 percent, with a particularly pronounced decline of 19 percentage points between exams 3 and 4. However, between those two exams, the individuals whose scores improved had relatively low average scores of 45 percent at exam 3 while those whose scores declined had higher average scores of 55 percent. Targeted feedback seems to have improved the scores of trainees who struggled to rate the staff-youth relations items accurately.

Average scores in the quality area of behavioral guidance were particularly volatile. After improving by 8 percentage points at exam 2, they fell by 23 percentage points at exam 3 and then rebounded to improve by 21 percentage points at exam 4. In the post-study survey, trainees frequently said that they disagreed with the master scores for behavioral guidance items. One trainee noted:

I think also I may disagree with some of the [master] scores in general. Although I understand we need to use the [master scores] as our guide, ... [w]hat the raters sometimes scored as inappropriate or disruptive behavior I felt was kids being kids.

Cultural differences in the interpretation of such factors as child behavior may have been responsible for some of the discrepancy in agreement scores among trainees from different racial and ethnic backgrounds.

Limitations, Implications, and Next Steps

We found a promising pattern of improved scores after trainees were exposed to the APT Anchors Guide and engaged in video-based practice. These findings suggest that future reliability trainings should focus on increasing familiarity with and expert knowledge of the guide. Ample opportunities for video practice should focus on improving accuracy by emphasizing links among ratings, particular events in the clips, and corresponding anchors.

This study has revealed that some APT items are more open to cultural-specific interpretations than others and that some videos are more ambiguous than others. Point of view is a key aspect of observational research, so there must be room for people from different cultural backgrounds to pose alternative interpretations of behaviors and to have different views of what constitutes low-quality and high-quality social interaction. We are

We found a promising pattern of improved scores after trainees were exposed to the APT Anchors Guide and engaged in video-based practice.

seeking additional funding to address any potential cultural bias in some APT items or training materials.

The study also exposed both the potential and the limitations of using video for training. Video technology is widely available and convenient, but problems such as connectivity issues can limit its usefulness. Furthermore, even with professional videographers using high-quality sound and video equipment, capturing the essence of youth-adult interactions is a tall order. Camera angles can provide limited views, and audio quality will vary depending on the size of the group and room. Early video practice did help participants modestly improve in their accuracy. Still, editing all videos to focus more carefully on the same visual and audio nuances may improve the ability of future video training to improve participant accuracy, particularly since people have limited attention spans for video viewing. The average viewing time for internet videos is only 2.7 minutes (Statistic Brain, 2016).

This article outlines the preliminary steps we have taken to test APT reliability training so that it can be further refined for wide adoption. Next steps include achieving an acceptable and consistent level of rater accuracy through video-based reliability training. An acceptable accuracy rating is usually set at 80 percent for similar tools in the field, such as the Center for Youth Program Quality's Youth Program Quality Assessment and TeachStone's Class (Bell et al., 2012). This prototype of a reliability training system with four exams must be further fine-tuned before going into the field for reliability certification. For instance, we would improve the system by taking into account the valuable trainee feedback, ranging from clarifying key terms in the APT Anchor Guide to carefully selecting video clips that are unambiguous.

The compelling reason to train observers to rate program quality accurately is that programs' use of such ratings is strongly associated with improving important quality areas such as supportive youth-staff relations and positive peer relations (Miller, 2005). As shown in the youth survey results in Table 2, these areas are significantly related to positive youth outcomes, such as sense of competence as a learner, sense of social competence, and future planning and expectations. Ultimately, we are refining the APT training so that it and the APT itself can be implemented more widely, where they can have an impact on youth program practice and policy and on the use of research evidence to support that critical work.

References

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta,

R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62–87.

Hill, H., Charalambous, C., McGinn, D., Blazar, D., Beisiegel, M., Humez, A., ... & Lynch, K. (2012, February). *The sensitivity of validity arguments for observational instruments: Evidence from the Mathematical Quality of Instruction Instrument*. Unpublished manuscript, Harvard University.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403–424.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43.

Lumley, T., & McNamara, T. F. (1993, August). *Rater characteristics and rater bias: Implications for training*. Paper presented at the Language Testing Research Colloquium, Cambridge, England. Retrieved from <http://files.eric.ed.gov/fulltext/ED365091.pdf>

Miller, B. M. (2005). *Pathways to success for youth: What counts in after-school*. Arlington, MA: Massachusetts AfterSchool Research Study.

Schlientz, M. D., Riley-Tillman, T. C., Briesch, A. M., Walcott, C. M., & Chafouleas, S. M. (2009). The impact of training on the accuracy of Direct Behavior Ratings (DBR). *School Psychology Quarterly, 24*(2), 73.

Statistic Brain. (2015, April 2). *Attention span statistics*. Retrieved from <http://www.statisticbrain.com/attention-span-statistics>