NEW FROM NIOST

Measuring Program Quality, Part 2

Addressing Potential Cultural Bias in a Rater Reliability Exam

Amanda Richer, Linda Charmaraman, and Ineke Ceder

Like instruments used in afterschool programs to assess children's social and emotional growth or to evaluate staff members' performance, instruments used to evaluate program quality should be free from bias. Practitioners and researchers alike want to know that assessment instruments, whatever their type or intent, treat all people fairly and do not privilege people from certain groups over others.

In the case of observation instruments, concern about bias extends beyond the instrument itself to the people doing the observation: how they apply the instrument's rubrics or standards in specific afterschool settings. A vital subset of concern about possible rater bias is whether any exam used to assess rater reliability itself carries unintended bias toward some groups of people.

This issue is not only a matter of fairness. Culturally biased rater reliability testing can directly affect youth outcomes. For example, an urban youth program in a low-income neighborhood with many people of color could receive negative scores from a rater who was not trained and certified to overcome any implicit biases related to racial and cultural practices different from his

AMANDA RICHER is a research associate and associate methodologist for the National Institute on Out-of-School Time and the Wellesley Centers for Women. She has been involved in psychometric testing of the APT since the first validation study and continues to use data to enhance OST tools.

LINDA CHARMARAMAN, PhD, is a research scientist at the Wellesley Centers for Women who has served as principal or co-principal investigator of two APT validation studies. She is engaged in using research evidence to inform policy and practice and in reducing structural inequality gaps in research.

INEKE CEDER is a research associate at the Wellesley Centers for Women whose work focuses on gender and race, women's leadership, and sex education. Her involvement on all three validity studies of the APT has driven home that research data can drive social change and bring equity for all.

or her own. When observation ratings affect funder decisions, the problem becomes acute.

Overcoming this possible source of bias is our concern in this article as members of the research and evaluation team for the Assessment of Program Practices Tool (APT). As we conducted studies to establish the scientific validity of APT (described in Tracy, Charmaraman, Ceder, Richer, & Surr, 2016), we uncovered apparent cultural bias in the preliminary results of the APT rater reliability exams: White raters tended to achieve the target rate of agreement with master scores more often than Black raters. This article describes the follow-up study we conducted to address the sources of that apparent bias, with the goal of making the APT rater reliability exam as free from cultural bias as possible. This goal is critical for any educational assessment, though it is often dismissed. During this follow-up study, we addressed practical concerns that have implications for the development of culturally fair program quality assessments across the field.

Culturally biased rater reliability testing can directly affect youth outcomes. For example, an urban youth program in a low-income neighborhood with many people of color could receive negative scores from a rater who was not trained and certified to overcome any implicit biases related to racial and cultural practices different from his or her own. When observation ratings affect funder decisions, the problem becomes acute.

Rater Reliability and Rater Bias

Inevitably, raters using observation tools are susceptible to their own biases (Hoyt & Kerns, 1999; Lumley & McNamara, 1993; Nisbett & Wilson, 1977). Hoyt (2000) argued that rater bias occurs when raters have their own personal interpretations of the measurement scale. Rater expectations also can be a source of bias (Rosenbaum, 2002).

Training and practice have been found to help minimize bias and increase rater accuracy (Chamberlain & Taylor, 2011; Hoyt & Kerns, 1999; Lyden et al., 1994; Schanche, Høstmark Nielsen, McCulough, Valen, & Mykletun, 2010; Schlientz, Riley-Tillman, Briesch, Walcott, & Chafouleas, 2009). Practice alone is not enough, but moderate to high dosages of training have been found to reduce rater bias (Hoyt & Kerns, 1999).

One strategy commonly used to achieve consistency and reduce bias is the use of explicit rating *anchors*. In an observation rubric, the anchors are detailed descriptions of what each point on the rating scale looks like, so that raters can clearly see what constitutes a rating of, for example, 1, 2, 3, or 4. Rater training that uses videos with a real-world example of each anchor has been shown to improve rater accuracy (Kishida & Kemp, 2010; Schlientz et al., 2009).

Another strategy to reduce bias is master scoring of video clips to establish a "gold standard" score. In this strategy, highly trained and experienced raters, usually working in groups of two or three, all rate the same videos, compare notes, and discuss until they can agree on a single master score for each video. Use of master-scored video training improves rater accuracy and mitigates rater "drift" (Bell et al., 2012; Hill et al., 2012).

The APT system uses these strategies—explicit anchors and master-scored videos—both in rater training and in the development of the rater reliability exams. When our validation study uncovered evidence of possible cultural bias in the results of the exams, we suspected that we had come up against an understudied yet crucial source of variance identified by Courtney Bell, senior researcher at Educational Testing Service (personal communication, June 6, 2016): that the master scores themselves had cultural biases that could unfairly privilege some groups of people.

The APT and Previous Validation Studies

The APT was launched by the National Institute on Out-of-School-Time (NIOST) in 2005 as an observation instrument to measure process quality: observable aspects of an out-of-school time (OST) program in action.

Designed to support program self-assessment and improvement, the APT is increasingly being used by external stakeholders across the country to ensure that afterschool programs are implementing quality features and to identify programs in need of improvement.

The APT has gone through three phases of reliability and validity checking (Tracy, Richer, & Charmaraman, 2016). *Reliability* is the extent to which an instrument produces consistent results; *validity* is the extent to which it measures what it is supposed to measure.

ABOUT THE APT

The APT measures process quality in three domains:

- Supportive social environment
- Opportunities for engagement in learning
- Program organization and structure

Each domain has subdomains called *quality areas*. For example, the quality areas for the domain *supportive social environment* are welcoming and inclusive environment, supportive staff-youth relationships, positive peer relationships, and relationships with families.

The items that measure these quality areas are spread throughout the APT, which is organized by program times of day: arrival, transition, homework, activity, informal, and pick-up.

The first APT validation study (APT I), funded by the William T. Grant Foundation and conducted with 25 OST programs in two states, aimed to establish reliability and to minimize measurement error. This study showed that the APT has many strong technical properties and is an appropriate tool for measuring afterschool program quality. However, it also found that rater reliability was somewhat unstable (Tracy, Surr, & Richer, 2012).

The purpose of the second validation study (APT II) in 2013–2015—again funded by the William T. Grant Foundation—was to develop and evaluate a multipronged reliability training. The training was designed to improve rater accuracy so that APT could be used for higher-stakes purposes, such as demonstrating program quality to funders. The data came from 39 rater participants from four states who completed reliability training including four online video-based exams. The training was improved from the previous iteration by expansion of the APT Anchors Guide, which explains the meaning of each possible score for each item; by video-based practice with immediate, detailed feedback; and by use of individualized reports that track rater progress in order to identify which video modules to focus on before the next exam. Accuracy scores improved slightly with these enhancements, but the average passing rates were still low, at 51 to 58 percent. The acceptable passing rate for similar tools in the field is 80 percent accuracy (Bell et al., 2012; Hill et al., 2012). The trainees provided valuable feedback on how to improve the training protocol, such as clarifying key terms in the anchors document and carefully selecting video clips that are unambiguous.

An unexpected finding was that Black participants had consistently lower accuracy scores than White trainees (see Charmaraman & Tracy, 2016). A follow-up analysis using logistic regression controlled for three aspects of compliance with the study protocol: consistent use of the APT Anchors Guide, the number of practice clips trainees rated, and watching the exam clips to the end. We still found significant differences in accuracy rates between Black and White and between Black and biracial participants, though all fell short of the 80 percent benchmark. This scoring gap between Black and White raters may be partially explained by the fact that Black raters, in qualitative feedback, often questioned the assigned master scores, rationales, and definitions. The feedback also suggested that use of shorter video clips would help raters achieve better accuracy by focusing their attention on specific instances.

The Current Validation Study

The primary goal of APT Validation Study III (2016–2017, funded by the William T. Grant Foundation) was to eliminate differences in accuracy rates between Black and White raters. We set out to identify sources of cultural bias, from the selection and narrative framing of the video clips to the assignment of master scores. The study had three research objectives:

1. To develop APT rater reliability exams in which the average rater score falls within the field benchmark of 80 percent
2. To refine APT rater reliability exams to reduce the potential for cultural bias and to examine whether demographic factors other than race or culture, such as gender, educational background, region of the country, number of years of OST experience, or experience with external program evaluations, are associated with better performance on the exams
3. To determine whether familiarity with the APT Anchors Guide, frequency of APT use, and APT training are positively associated with better performance on the rater reliability exams

Training and Exam Materials

We built on the work of the first two studies to refine the selection of videos to use for training and for the rater reliability exams. We also modified the language in the APT Anchors Guide and set up a process to produce new master scores for the selected videos.

To develop the training and exams for APT II, we had videotaped activities at eight programs in New England. For the current study, we reassessed the library of video clips to reduce confounding factors such as ambiguous elements or extraneous details, including issues with the length or quality of the clip. We selected video clips that focused only on one of the six APT time-of-day sections—the Activity section—and on elementary (grades K–5) programs. Three experts in the development and use of the APT selected 35 clips that met these criteria. The three experts had to agree on which subscales of the APT Activity section (see box) the clips exemplified. One clip might be rated on three to five items within those subscales. The clips varied in length from 1 minute to 8.5 minutes, with most hovering around 3 to 4 minutes. For use in practice sessions and rater reliability exams, each clip was preceded by a short description of what was taking place; whether the clip showed the beginning, middle, or end of the activity; and which subscales from the Activity section the participant would be rating.

We reviewed the APT II version of the APT Anchor Guide for potential ambiguities. To reduce ambiguity in anchor descriptions, whenever possible we added quantities of how much something occurs or how many people participate. Guided by our analytical results from APT II, we formed a working group comprising the authors of this article and NIOST staff to identify culturally ambiguous items and to reduce ambiguity by producing more descriptive definitions and examples of phrases like "inappropriate behavior" that might have different meanings for different groups of people. In order to reduce variation in how often raters referred to the APT anchors while rating clips, we included the anchors in the practice modules and exams themselves, rather than providing a separate guidebook.

APT ACTIVITY SECTION SUBSCALES

- Organization of activity
- Nature of activity
- Staff promote youth engagement and stimulate thinking
- Staff positively guide youth behavior
- Staff build relationships and support individual youth
- Youth relations with adults
- Youth participation in activity time
- Peer relations

Master scores for APT II were provided by predominantly White raters. For this third study, we recruited four consultants of color to serve on the master scoring team. All had extensive expertise as afterschool directors, as evaluators, or as APT trainers. Two of them had participated as master scorers in APT II. All four consultants were female; three were African American, and one was Latina.

Before they rated the 35 selected video clips, we required the master scorers to review a document to sensitize them to cultural bias. We gave them the revised APT anchors and shared feedback from Black participants in APT II who disagreed with master scores for clips. After reviewing these materials, each consultant rated each video clip. If three of the four agreed on a rating for an item, then that became the master score. If not, then a fifth consultant from the previous APT master scoring committee, a White male, served as "tiebreaker." If three of the five agreed, then that became the master score. If not, the clip was discussed at a consensus meeting. All five group members then had to agree for the clip to be included in this study. We recorded the reasons these consultants gave for their ratings and used these reasons to develop practice materials.

Pilot and Field Testing

We sent the APT Anchors Guide to a total of 16 pilot participants, 30 percent of whom were non-White, and asked them to get familiar with it. A few days later, we sent them an email with links to three practice clips and three exams. These consisted of short video clips, each followed by the APT Activity subscales, such as *organization of activity* and *youth relations with adults,* on which participants were to rate the clip using the APT anchors. During the pilot tests, participants could share feedback on, for example, whether the clip was connected to the right APT scale, whether it showed enough information to enable them to rate it properly, and whether audio and video were of high enough quality. Feedback enabled us to fine-tune the final version of the three exams. For example, for the ensuing field test, we displayed the specific APT subscales to be rated before showing each video clip so raters would know what they were looking for.

After the pilot tests, we recruited 32 field-test participants, who also completed three online practice sessions and three exams. Participants were instructed to rate the practice clips first, before they began taking the three exams. For the practice clips, they received feedback on the accuracy of their ratings and were shown the reasons the master scorers had given for their ratings. For the exams,

they received only feedback on their accuracy but not the rationales for the master scores. The order in which participants took the practice clips and exams was randomly assigned to prevent any measurement error from the "order effect," in which the order of the exams can significantly affect the results.

Study Participants

To select participants for APT III, we tapped a database of APT users trained directly or indirectly by NIOST within the last 10 years, inviting 537 individuals to field-test three APT rater reliability exams. The email invitation included a short survey to gather information about demographics and APT experience. Of the 537 candidates, 97 responded by filling out the demographic survey; of those, 48 (49.5 percent) ended up participating in the study.

The 48 participants came from 11 states, and 33 percent were non-White. This sample thus was more diverse in geography and race or ethnicity than those of previous APT validation studies. Table 1 outlines the demographic characteristics of the sample. In terms of experience, a substantial proportion, 87 percent, had experience with K–5 students; many reported working with students through grade 12. Most participants were familiar with the APT anchors (73 percent) and almost half reported using the APT one or two times per year. Asked about APT training, 79 percent reported having received in-person NIOST training, 52 percent online NIOST training, 25 percent training at their own site, and 27 percent training in a previous APT validation study. Participants could report having received more than one type of training.

Data Analysis

The final analysis sample combined exam data from the pilot tests and field tests with a total of 48 participants.

Item-Level Analysis

Following advice we solicited from expert methodologists, we explored the range of scores for each exam item and compared participants' ratings to the master scores. The goal of item-level analysis was to create exams that would be practical for use in the field. For most items, a majority of raters exactly matched the master scores. However, a few items on each exam had poor accuracy rates, typically less than 40 percent; also, the variation in scores was more than just one point on the four-point scale. For some conditional items, where raters would need to see a particular condition—for example, children behaving inappropriately—in order to rate the item, many participants considered the condition to have occurred while others did not. For these reasons, a few items were removed from each exam.

Many other items were assigned two accurate ratings. Other observation scales in the field, such as Teachstone's CLASS instrument (Bell et al., 2012), consider a rating to be accurate when it falls within one point of the master code. The decision to allow two accurate ratings addresses the issue of assigning one "accurate" quantita-

Table 1. Demographic Characteristics of Study Participants

Characteristic	Percentage of Sample ($N = 48$)
Female	77%
White	67%
Black	17%
Hispanic	10%
Asian	4%
Native American	2%
Age 20–29	19%
Age 30–39	38%
Age 40+	44%
Work in the Northeast	73%
Work in the South	21%
Work in the West or Midwest	6%

tive score to qualitative observational ratings, which are subject to personal biases. It allows for the possibility that the "true" rating could land in between two scores. For APT rater reliability exams, most items required two accurate scores. The stringent criterion for assigning a single master score to an item—that one consistent best score was assigned by raters across most groups, so that one group was not unfairly privileged over another—was met by 35 percent of the items.

Rater-Level Analysis

To assess rater accuracy, each participant's score for each item was compared to the master score. A rater accuracy score was calculated for each participant by dividing the total number of items rated correctly by the total number of items in all three exams. Using the rater accuracy score, a percentage correct score was calculated for each participant for each exam. Statistical tests were used to assess group differences in rater accuracy scores.

Findings

We report our results under headings related to the three research objectives: rater reliability, group differences, and the effects of APT experience.

Rater Reliability

The first research objective was to reach average rater accuracy scores that fell within the field benchmark of 80 percent.

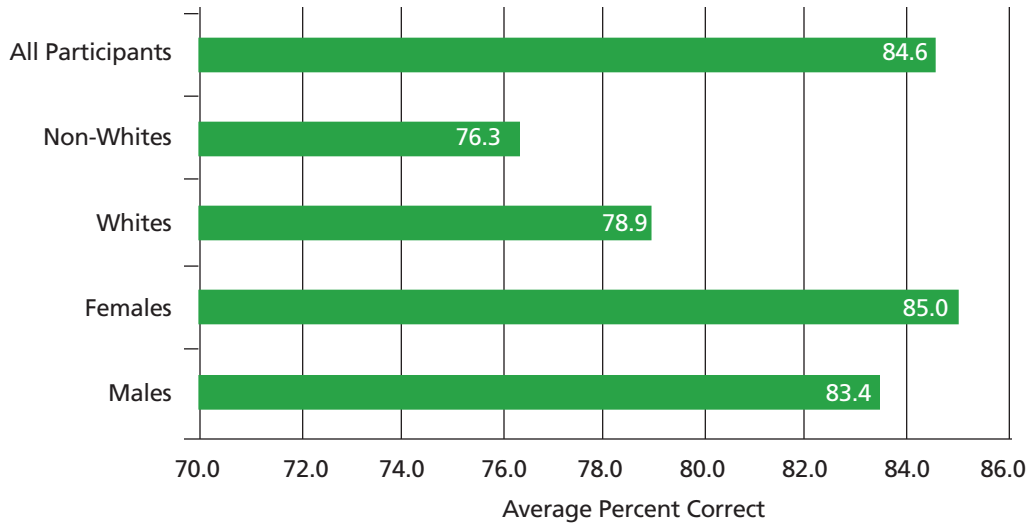
As was the case in APT II, average rater accuracy scores were initially lower than the field benchmark of 80 percent: 58.8 percent for Exam 1, 57.2 percent for Exam 2, and 61.4 percent for Exam 3. When we removed problematic items from the exams and allowed items to have two correct answers, the average rater accuracy scores increased to 82.4 percent for Exam 1, 84.9 percent for Exam 2, and 86.5 percent for Exam 3. The rate at which raters passed the benchmark of 80 percent was also calculated for each exam. The analyses exploring group differences used these rater accuracy scores and benchmark passing rates to test statistically for group differences among raters.

Group Differences

The second research objective was to examine differences in rater accuracy scores to look for group-level biases by demographic categories and by experience in OST programs.

We conducted group difference tests by gender, race, age, region, and education background on the average rater accuracy scores and benchmark passing rates. Figure 1 shows average rater accuracy rates for selected demographic characteristics. No significant differences were found between males and females, White and non-White participants, or people residing in and outside of New England (where the APT was developed and videos were recorded); nor were there differences among age groups. For educational background, we found a significant group difference for one exam only, showing that

Figure 1. Average Rater Accuracy Scores by Gender and Race



participants with a PhD were less accurate compared to those with a bachelor's or master's degree. In relation to these demographic categories overall, we found no significant group-level biases in average rater accuracy scores in relation to benchmark passing rates, indicating that the reliability exams do not favor one type of rater over another.

We also looked for differences based on participants' OST experiences. One variable was experience with different age groups. We looked for differences between participants with K–5 experience and those with none and for differences among participants with K–5 experience only, K–8 experience only, and K–12 experience. We found no differences in average rater accuracy scores in relation to grade-level experience. Nor did we find differences for participants who reported having worked with minority students, with low-income students, with students in urban environments, or in large programs with high student-to-staff ratios, as compared to participants who did not report having these experiences. These results suggest that the exams demonstrate no bias toward raters who have worked with K–5 students (the ones depicted in the videos) or toward those who have or have not worked with vulnerable OST populations.

The Effect of APT Experience

The third research objective was to discover whether familiarity with the APT Anchors Guide, use of APT in the field, or APT training led to better performance on the rater reliability exams.

The only factor that had a significant effect on accuracy was familiarity with the APT anchors. For all three exams, raters who were familiar with the APT anchors were more likely to pass the exam at the 80 percent benchmark than those who not were familiar with the anchors. For two of the three exams, raters familiar with the anchors also had higher total accuracy scores.

For frequency of APT usage, only one exam showed a difference in accuracy between raters who used the APT three to five times per year and those who used it five or more times per year. Similarly, when we looked for differences among raters who had and had not used the APT for external evaluation, we found significant differences in one exam.

For the effect of APT training, results were mixed. Participants reported what type of APT training they had received throughout their experience and how long ago they had received this type of training. We found significant differences by the number of types of training participants had experienced for one of the three exams.

Implications

In this study, the APT rater reliability exams achieved rater accuracy levels meeting the benchmark passing rate of 80 percent. We found no significant group differences in rater accuracy among the three exams, suggesting that they are equivalent. We found no significant differences among raters by race, gender, age, region, or experience with OST populations.

Ample evidence demonstrates that familiarity with the APT anchors is associated with higher rater accuracy. Our findings also suggest that frequency and type of APT use may have some relationship to rater accuracy. This relationship, along with the relationship between APT training and rater accuracy, warrants further investigation. APT training is a prerequisite for knowledge of the APT anchors and for use of the APT anchors. The relationship between training and rater accuracy therefore needs further evaluation with a larger sample. Development and evaluation of specialized APT training focused on improving rating reliability would be the next step.

The finding that familiarity with the APT anchors improves raters' ability to pass the reliability exams is key to our goal of creating exams that treat all groups fairly. A malleable intervention, such as improving familiarity with the APT anchors, may be what drives accuracy levels, rather than any static demographic trait such as race.

Our process and findings suggest practical implications for rater reliability testing in two interrelated areas: use of master scores and steps to reduce cultural bias.

Use of Master Scores

As we conducted this study, we explored the advantages and disadvantages of using master scores, in which a group of expert raters assigns one correct score to each item on a reliability exam. The advantages of master scoring are that it:

- Standardizes ratings and rating accuracy across programs and sites
- Reduces the effect of internal raters' bias stemming from familiarity with the program and its staff
- Improves raters' awareness of the need for objective evidence and descriptive examples to justify ratings

Disadvantages of master scoring to establish one "best" score include the following:

- Inherent problems with the idea that there can be only one "best" score for each item
- The false expectation that a single less experienced rater could arrive at the same score as a group of expert raters
- Inability to allow for two "best" scores when many raters believe an item falls between scores

Extensive discussion with methodologists in the field convinced us that one master score may *not* be the only score that is true and accurate. In real-world observations, raters often find themselves wanting to rate "in the middle" between two ratings—for example, the score is not 2 or 3 but 2.5. Another important consideration is that the expert raters who produced the master scores did not do so in isolation. They often disagreed on ratings for individual items (or wanted to rate them "in between") and needed the group process of master consensus meetings to reach 100 percent agreement. Individuals taking a rater reliability exam—or rating program quality in the field—do not have access to such a group process. Expecting a single less experienced rater to consistently arrive at the same score as a group of highly experienced raters is simply unrealistic.

These considerations led us to identify items that had strong leanings toward two possible scores. Allowing two scores for a single item helps to compensate for limitations in the video clip itself, such as length, sound quality, or camera viewpoint, that could produce ambiguity. More importantly for the purpose of this article, allowing two scores also accommodates different cultural and contextual interpretations by raters from a wide variety of backgrounds.

> The finding that familiarity with the APT anchors improves raters' ability to pass the reliability exams is key to our goal of creating exams that treat all groups fairly. A malleable intervention, such as improving familiarity with the APT anchors, may be what drives accuracy levels, rather than any static demographic trait such as race.

Steps to Reduce Cultural Bias

In the process of refining the video-based APT rater reliability exams to reduce the potential of cultural bias, we:

- Selected video clips with as little cultural ambiguity as possible, so that they would be less prone to different interpretations by raters from different cultural backgrounds
- Selected a racially diverse panel of master scorers
- Provided those master scorers with cultural bias training
- Revised the APT Anchors Guide to define key terms that could be read differently by people from different backgrounds

As we worked to eliminate cultural bias in the APT rater reliability exams, we developed a checklist of categories that are often subject to cultural bias during program quality observation, including socioeconomic status, urbanicity, program size, racial and ethnic backgrounds of students and staff, gender, and what constitutes "appropriate behavior" in different cultures. The people who know best which of these factors are at play in a given program setting are not external observers but program directors and staff. We therefore strongly suggest that program directors and raters—before, during, and after program quality assessments—become aware of and attempt to address potential biases. For instance, in the cultural bias training, we ask master scorers to pay attention to biases related to socioeconomic status. We ask them to reflect, for example, on whether they are giving higher ratings to programs with high-quality materials and activities that cost more while unintentionally assigning systematically lower ratings to programs with smaller budgets.

Policy Implications

Our study is a contribution to ongoing discussion in the OST field about cultural bias in program quality assessment. In order to make smart decisions about effective educational interventions and resource allocation, the OST field needs evidence from research. To provide accurate and reliable evidence, researchers must develop—and funders and policymakers must seek and support—assessments that reduce scoring gaps favoring one group over another. Culturally informed test development practices can affect how programs and staff members are supported. When funding decisions depend on the results of program quality assessments, cultural bias in those assessments can have a direct effect on program youth. To be fair to youth, their families, and their communities, the field needs culturally fair assessments of program quality.

Acknowledgements

This research was generously funded by Officer's Research Grant #187010 from the William T. Grant Foundation. Publication support was also provided to Linda Charmaraman by New Connections: Increasing Diversity of RWJF Programming at the Robert Wood Johnson

Foundation. We would like to thank Lisette DeSouza for her contributions and our study participants for making this research possible.

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.

Chamberlain, S., & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology*, *42*(4), 665–675.

Charmaraman, L., & Tracy, A. (2016, August). *Avoiding cultural bias in establishing program observation accuracy: Reflections on our evolving action-oriented mixed methods evaluation design*. Paper presented at the Mixed Methods International Research Association conference, Durham, United Kingdom.

Hill, H., Charalambous, C., McGinn, D., Blazar, D., Beisiegel, M., Humez, A., Kraft, M., Litke, E., & Lynch, K. (2012). The sensitivity of validity arguments for observational instruments: Evidence from the Mathematical Quality of Instruction instrument. *Educational Assessment*, *17,* 1–19.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64–86.

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

Kishida, Y., & Kemp, C. (2010). Training staff to measure the engagement of children with disabilities in inclusive childcare centers. *International Journal of Disability, Development and Education,* 57(1), 21–41.

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.

Lyden, P., Brott, T., Tilley, B., Welch, K. M., Mascha, E. J., Levine, S., Haley, E. C., Grotta, J., & Marler, J. (1994). Improved reliability of the NIH Stroke Scale using video training. NINDS TPA Stroke Study Group. *Stroke, 25*, 2220–2226.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250–256.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Schanche, E., Høstmark Nielsen, G., McCulough, L., Valen, J., & Mykletun, A. (2010). Training graduate students as raters in psychotherapy process research: Reliability of ratings with the Achievement of Therapeutic Objectives Scale (ATOS). *Nordic Psychology*, *62*(3), 4–20. doi:10.1027/1901-2276/a000013

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), 73–83.

Tracy, A., Charmaraman, L., Ceder, I., Richer, A., & Surr, W. (2016). Measuring program quality: Evidence of the scientific validity of the Assessment of Program Practices Tool. *Afterschool Matters, 24*, 3–11.

Tracy, A., Richer, A., & Charmaraman, L. (2016, April). *APT validity and reliability: Identifying and minimizing measurement error of youth program observation ratings*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

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.