Published in final edited form as:

J Early Adolesc. 2015 ; 35(5-6): 681–713. doi:10.1177/0272431614561261.

Measuring afterschool program quality using setting-level observational approaches

Yoonkyung Oh, D. Wayne Osgood, and Emilie Phillips Smith

The Pennsylvania State University

Abstract

As the importance of afterschool hours for youth development is widely acknowledged, afterschool settings have recently received increasing attention as an important venue for youth interventions. A range of intervention programs have been in place, generally aiming at positive youth development through enhancing the quality of programs. A growing need has thus arisen for reliable and valid measures of afterschool quality. This study examined the extent to which the two observational tools, i.e., Caregiver Interaction Scales (CIS) and Promising Practices Rating Scales (PPRS), could serve as reliable and valid tools for assessing the various dimensions of afterschool setting quality. The study shows the potential promise of the instruments, on the one hand, and suggests future directions for improvement of measurement design and development of the field, on the other hand. In particular, our findings suggest the importance of addressing the effect of day-to-day fluctuations in observed afterschool quality.

Keywords

Caregiver Interaction Scale; Promising Practices Rating Scale; Afterschool Program Assessment; Reliability; Validity; Generalizability theory

After-school programs have become an important setting for aiding schools' mission of promoting positive youth development and academic achievement, and accumulating evidence indicates that program quality is critical to their success (Mahoney, Parente, & Lord, 2007; Pierce, Bolt, & Vandell, 2010). Reliable and valid measures of program quality are essential both for research seeking to enhance the effectiveness of after-school programs and for assessments to ensure the quality of practice. We present a study of the measurement properties of two observational measures of the quality of after-school programs. One is well-established, and we seek to determine its applicability to a broad range of populations and locales. The other measure, newly applied to the after-school context, is of interest both because it addresses one aspect of program quality in greater depth and because it may be useful for assessing important aspects of program functioning other than overall quality (Shinn and Rapkin, 2000). The study not only addresses standard psychometric features of these measures, but it also applies generalizability theory to assess the separate and joint contributions of observers and measurement occasions to reliability measures (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Raudenbush and Sampson, 1999).

Pre- and early adolescence is a developmental period with heightened risk of problem behavior as well as great opportunity for promoting positive development. Preventive interventions during this period are important because earlier experiences can shape long-term developmental trajectories (Eccles & Roeser, 2009). Such interventions are often delivered in settings such as family, school, and community, which theorists commonly acknowledge as major influences on development (Bronfenbrenner, 1979; Gorman-Smith, Tolan, & Henry, 2000). Afterschool settings have received increasing attention as an important venue for supporting schools' missions to promote both academic achievement and positive youth development (Durlak, Weissberg, & Pachan, 2010; Eccles & Gootman, 2002; Kuperminc, Smith, & Henrich, 2013). These contributions from afterschool settings have grown as school settings face pressure to raise performance on standardized tests, which leaves less opportunity for non-academic activities (Mahoney & Zigler, 2006).

Accumulating research has shown that high quality afterschool settings make a significant difference in outcomes such as social skills, motivation, and academic performance (Mahoney et al., 2007; Pierce et al., 2010; Pierce, Hamm, & Vandell, 1999; Riggs, 2006). For instance, persuasive evidence demonstrates that participating in out-of-school offerings such as afterschool programs has a substantial impact upon teacher-reported academic, social, and prosocial skills of both elementary and middle school students, with effect sizes for elementary students ranging from .21 to .51 (Vandell, Reisner, & Pierce, 2007). This research has spawned a growing number of efforts to improve process and practice in afterschool programs (e.g., Durlak et al., 2010; Gortmaker et al., 2012).

Pianta and his colleagues have argued that validated and objective assessment of classroom processes is the foundation for programs, interventions, and policies to improve educational services (Hamre, Pianta, Mashburn, & Downer, 2007; Pianta, 2005), and we believe this point applies to afterschool settings as well. However, few systematic, reliable, and well-validated methods exist for measuring the quality of afterschool settings. Compared to a long record of measurement tools developed for school, classroom and early child care settings (for reviews, see Halle, Vick Whittaker, & Anderson, 2010; Pianta & Hamre, 2009), research on afterschool quality measurement is still in its early stage (See Yohalem & Wilson-Ahlstrom, 2010 for a review).

Our work concerns measures of afterschool program quality. Although research supports the benefits of attending afterschool programs (Durlak et al., 2010; Lauer et al., 2006), positive effects are not observed in all afterschool programs. Some programs help to promote positive youth development while other programs provide more opportunities for antisocial involvement and fewer opportunities for skill building (Mahoney & Stattin, 2000; Mahoney, Stattin, & Lord, 2004), leaving open the question of what constitutes quality afterschool programs. Our research seeks to contribute to the development of reliable and valid measures necessary for research seeking to establish which dimensions contribute most to program effectiveness and which program practices best enhance the dimensions of quality.

Our study examines two existing observation instruments in a sample of afterschool programs serving kindergarten up to middle school students. We evaluated the Promising Practices Rating Scale (PPRS: Vandell et al., 2004) developed specifically to measure the

quality of afterschool programs. We also tested the Caregiver Interaction Scale (CIS: Arnett, 1989), which was designed to measure the quality of early child care settings. Because the CIS provides more detailed information about staff-student interactions than existing measures of afterschool settings, we examined its applicability in a context attended by older children and youth. In the next section, we review the salient domains of afterschool program quality that are known to affect youth development and that these instruments intend to measure. We then review several observational approaches to assessing the quality of afterschool programs, explain the ways in which our study advances beyond prior research, and then proceed to our empirical work.

Key Dimensions of Afterschool Quality

Supportive relationships with staff

A broad literature documents that supportive relationships with extra-familial adults help to reduce problem behaviors and improve school-related outcomes (DuBois, Holloway, Valentine, & Cooper, 2002). Positive relationships with teachers have been conceptualized as a primary mechanism through which classroom experiences lead to positive youth outcomes (Pianta & Allen, 2009). Similarly, in afterschool research, supportive relationships with staff differentiate high-functioning afterschool programs from less effective ones (Eccles & Gootman, 2002; Grossman, Goldmith, Sheldon, & Arbreton, 2009). Specifically, supportive relationships with afterschool staff are related to increased learning and engagement in afterschool activities (Grossman et al., 2009); heightened sense of self-efficacy (Walker & Arbreton, 2004); fewer internalizing and externalizing problems (Pierce et al., 1999); and higher reading and math grades (Pierce et al., 2010).

Supportive relationships with peers

The accumulated body of research indicates that academic, psychological, social and behavioral adjustment of children and youth is influenced by peer relations, whether measured as characteristics of peers or peer groups (Véronneau & Dishion, 2011), peer attachment and peer support (Gonzales, Cauce, Friedman, & Mason, 1996), friendship quality (Franco & Levitt, 1998) or peer norms (Shin, Daly, & Vera, 2007). Furthermore, Smith and colleagues (in press) have shown that though supports from adult family members and the community are associated with reduced delinquency, positive peer interactions are more strongly related to positive youth development. Research evidence indicates that positive peer relations in afterschool programs are one of the primary mechanisms through which afterschool program participation affects a wide range of youth outcomes, including delinquent behavior (Gottfredson, Gerstenblith, Soulé, Womer, & Lu, 2004); internalizing and externalizing problems and social skills (Pierce et al., 1999); and sense of self-efficacy and response to social challenges (Walker & Arbreton, 2004).

Developmentally appropriate structure and supervision

Provision of structured and supervised time after school is the core *raison d'être* of afterschool programs (Eccles & Gootman, 2002), and unstructured socializing with peers in the absence of adult supervision increases delinquency and problem behaviors among youth and adolescents (Maimon & Browning, 2010; Osgood & Anderson, 2004). Mahoney and

colleagues have found that participation in unstructured and less-well supervised programs increases rather than decreases youth involvement in antisocial behavior (Mahoney & Stattin, 2000; Mahoney et al., 2004) while Gerstenblith and colleagues (2005) have shown that youth benefit from attending structured afterschool programs with organized transitions, planned activities, structured skill-oriented programming, and quality behavior management.

Youth engagement

Opportunities for youth to make a difference or “matter” in their environments are associated with positive outcomes (Eccles & Gootman, 2002). McLaughlin, Irby, and Langman (1994) found that community-based programs with high levels of youth participation provided increased opportunities to build skills, competencies, and self-worth. Highly engaging activities offer opportunities that promote achievement, skill-building, and intrinsic motivation (Larson, 2000). Furthermore, youth involved in interesting programs are less likely to discontinue participation in the program (Perkins & Borden, 2003). Higher levels of engagement in afterschool programs are linked with positive youth outcomes, including increased social competence (Mahoney et al., 2007; Shernoff, 2010), improved intrinsic motivation (Mahoney et al., 2007), and decreased school suspension rates (Blazevski & Smith, 2007).

Measuring Quality in Afterschool Settings

We focus on observational measures of the setting-level quality of afterschool programs. Administrative data and surveys also could be considered for the assessment of setting quality. Administrative data offer a convenient way to collect information on compositional characteristics and features such as locale and size, but they do not capture the social processes (Raudenbush & Sampson, 1999) that intervention programs seek to change. Surveys can be used to collect information on social processes and aggregated to produce setting-level measures. This method has the advantage that setting quality is evaluated through the lenses of individual participants (i.e., staff and children) who know the setting well (Fraser, 1998), which is especially valuable for some purposes. However, the setting-level aggregates of individual perceptions may be colored by individual beliefs, histories, values, desires and expectations (Kirtland et al., 2003), which may be problematic for other purposes. In fact, individuals’ perceptions of the same setting typically show a low level of agreement, which limits the reliability of survey-based setting-level measures (Raudenbush & Sampson, 1999).

Direct observational measures provide information from an uninvolved observer, who presumably has no personal interest in whether any given program receives higher scores than another (Nock & Kurtz, 2005). Observational measures are, of course, subject to their own sources of error, including observer drift over time, observer bias stemming from prejudices and expectations about the study and the subjects, and the reactivity of the subjects to being observed (McCall, 1984). Unlike survey-based methods, however, researchers can reduce observer errors by careful recruitment of observers, clear specification of observer roles, and adequate and continuous training (McCall, 1984; Skinner, Dittmer, & Howell, 2000).

The Present Study

Yohalem and Wilson-Ahlstrom's (2010) review of ten measures of after-school program quality demonstrated that scholars have invested considerable effort in developing such measures. We seek to advance these efforts by examining the measurement properties of two instruments. We chose the Promising Practices Rating Scale (PPRS, Vandell et al., 2004) because it best matches the dimensions of quality we consider important (discussed above), among the measures with previous evidence of reliability and validity. Other options would have been the School-Age Child Environment Rating Scale (SACERS, Harms, Jacobs, & White, 1996), and the Youth Program Quality Assessment (YPQA, C. Smith & Hohmann, 2005). The SACERS was somewhat less suitable to our aims because it does not allow us to differentiate between supports from various sources, including teachers, family, and peers. The YPQA would have been a reasonable choice for our investigation as well, but its reliability and validity have been addressed more extensively elsewhere than have the measures on which we focus.

The PPRS is especially well-suited to our interests because it includes separate scales for supportive relationships with adults, supportive relationships with peers, appropriate structure, level of engagement, and chaos (a problematic manifestation of inadequate structure).¹ In prior research, the PPRS has exhibited correlations with youth school-day academic performance, particularly among urban, ethnic-minority populations in the U.S. (Vandell, Reisner, & Pierce, 2007). More information on youth who vary in race-ethnicity and geographic locale would be informative about the usefulness of this tool in assessing out-of-school time care quality.

The second measure we investigate, the Caregiver Interaction Scale (CIS, Arnett, 1989), previously has not been used in after-school settings. The CIS is of special interest for differentiating several aspects of one of the most central domains of program quality, namely, interactions of staff with students. More specifically, the CIS includes subscales for staff harshness, permissiveness, detachment, and sensitivity. This instrument was designed to measure early child-care settings, and we examine its applicability and utility in afterschool settings attended by older children and youth. Though the present research is limited to setting level assessments, a distinctive appeal of this scale is that it provides data useful for a broader range of purposes. Because the CIS examines the practices of individual staff, it would facilitate research that examines not only overall scores for programs, but also individual staff ratings, aggregated staff ratings, and the degree to which there is dispersion and variation among the staff (Shinn & Rapkin, 2000). Therefore, we believe it is worthwhile to assess whether our adaptation of the CIS has adequate measurement properties when used to assess afterschool programs for elementary and middle school students.

Our analyses of the measurement properties of the CIS and PPRS include assessments of internal consistency, inter-rater reliability, and dimensional structure. To determine whether

¹The full PPRS also includes subscales for Over-control, Mastery, and Opportunities for Cognitive Growth, which are not included in the present study. Our data come from the evaluation of an intervention that did not seek to affect those, so they were omitted from the data collection.

measurement properties are consistent across types of programs, additional analyses compare afterschool programs that predominantly serve low income, ethnic minority, urban youth with those that do not. In light of wide variations across afterschool settings and challenges of serving predominantly disadvantaged populations, it is important to consider whether afterschool program quality can be measured similarly across different settings.

In addition to these standard psychometric topics, we address special issues for observational measures of settings that have received little attention for afterschool programs (Raudenbush and Sampson, 1999). Not only are observations subject to disagreement between observers, but also to day-to-day fluctuations in program functioning. The greater this variability, the more precision will depend on the amount of observation time. Estimates of variance due to these sources of error indicate the benefits of increasing the number of observers and occasions per site when designing observation systems.

Method

Participants

This study is part of LEGACY Together², which is a randomized controlled trial study of a behavioral intervention in afterschool settings. The project involved 1,200 children and staff and 75 afterschool programs in Pennsylvania. Data were collected from three cohorts over a three-year period (2009–2012) using data from various sources (children, staff, director, school teachers) and methods (direct observation, self-report surveys, documents).

The present study used data drawn from the observation of 44 afterschool programs that participated in the first two years of the larger program evaluation project. The programs were clustered in a limited set of geographic locations in central or eastern Pennsylvania, and inclusion required agreement to participate in a random trial of a program to enhance effectiveness. These afterschool programs were diverse in mission, locale and demographic composition, which is a valuable strength for a measurement study. All of the programs were run by community-based organizations and housed in school buildings. The programs included 2 21st Century Community Learning Centers, 23 programs operated by YMCAs in various communities, and 19 programs operated by private providers, with many of the latter participating in the *Keystone STARS program*—a quality rating system for improving childcare settings in Pennsylvania. The primary activities of all the programs were snacks, homework, and recreation. Nine of the programs put greater emphasis on structured learning activities than did the others, but none was predominantly academic. The programs typically served youth from the end of the school day until 5:30 or 6:00. All operated five days per school week, with most (but far from all) students attending almost every day.

The demographic characteristics of the program participants appear in Table 1. The locales of the afterschool sites were 11% rural, 57% suburban, and 32% urban. In terms of racial/ethnic composition, 36% were predominantly white (minority enrollment < 25%) and 27% of the programs had a high concentration of minority children (minority enrollment > 75%). The programs were also diverse in socioeconomic composition: 25% of the programs had

²LEGACY (*Leading, Educating, Guiding A Community of Youth*) Together

disproportionately high percentages of low-income children with over 75% eligible for free or reduced price meals while another 25% were economically advantaged with less than 25 % eligible. These afterschool programs were small to moderate in size: the number of enrolled children ranged from 11 to 50 ($M=24.24$, $SD=10.67$) and the number of staff ranged from 1 to 8 ($M=3.75$, $SD=1.75$). All the programs enrolled children between the grades of kindergarten and 7th ($M=2.5$, $SD=1.70$).

Direct Observation Tools

Arnett's Caretaker Interaction Scale (CIS)—The CIS was used to measure the quality of staff-child interactions in our afterschool programs. Developed by Arnett (1989), the original CIS consisted of 26 items examining the interactions of caregiving staff with preschool children in the following 4 areas: a) harshness and more authoritarian manner, expecting children to rigidly conform to adult rules and behavior; b) sensitivity, which is defined as not only communication with young people but the provision of appropriate levels of guidance, redirection, praise, and contingent rewards; c) detachment, uninterested in children and involved in more adult-oriented activities that exclude interaction with children; and d) permissiveness, the degree to which staff fail to appropriately provide guidance and redirection when necessary. Based upon our pilot work, we adapted some of the items to make them more observable and as unambiguous as possible, resulting in 23 items (see Table 2 for the full list of items). Trained field observers rated up to 3 permanent, non-volunteer staff³ in each afterschool program on a 4-point response scale indicating the extent to which they engage in a particular behavior or practice : 1=*never* (0%); 2=*few instances* (1~30%); 3=*many instances* (31~60%); and 4=*consistently* (>61%).

Promising Practices Rating Scale (PPRS)—The PPRS is an observational tool developed by Vandell and colleagues (2004) to study the relationship between afterschool program quality and youth outcomes. The PPRS was designed to collect information on types of activities offered in afterschool programs, resources available for those activities, and various dimensions of program quality and practices during those activities. These dimensions included: supportive relations with adults, supportive relations with peers, level of engagement, opportunities for cognitive growth, appropriate program structure, over-control, chaos, and mastery orientation.

We modified the original version of the PPRS in several ways. In the original PPRS, each dimension of program quality was assessed by a single global item with a 4-point scale indicating the extent to which a given construct is characteristic of the program (1=*highly uncharacteristic*; 2=*somewhat uncharacteristic*; 3=*somewhat characteristic*; and 4=*highly characteristic*). Single-item measures, however, may have limited reliability (Nunnally & Bernstein, 1994). In order to address this concern, we developed multi-item subscales representing each dimension of program quality by treating specific exemplars for each dimension provided in the PPRS as separate items that could serve as subscale components. If the PPRS manual included more than four exemplars, then four were chosen to represent

³Permanent staff included both full-time and part-time staff. Part-time staff comprised the majority of the afterschool workforce (75% or more) in 80% of the programs. When more than 3 staff were eligible for observation, observers selected staff who engaged with students more directly.

as much as possible the full range of content in the entire set.⁴ The ratings on the set of items for each dimension were then averaged to obtain a subscale score.

Next, we modified the standard PPRS observational protocol, which calls for the observation of 6 successive activities for 15-minute intervals (Vandell et al., 2004). Our data collectors instead observed 3 successive activities⁵ during each day of observation for an afterschool program. Doing so reduced the required time, which was necessary for gathering other data. Compensating for this reduced sample of activities at each visit, we observed each program twice over a period of four to six weeks. This provided an equal amount of observing time while also enabling us to determine the contribution of day-to-day variation to measurement error.

Another modification was to omit three of the PPRS subscales, over-control, mastery orientation, and opportunities for cognitive growth because the program being evaluated did not seek to affect these dimensions. Our implementation of the PPRS scale therefore consisted of 17 items, with the following five subscales: supportive relations with adults, supportive relations with peers, level of engagement, appropriate structure and chaos. (See Table 3 for the list of items).

Data Collection Procedure

Direct observation is a complex and costly approach to setting-level measurement, but likely provides the least biased assessment of social ecological settings. As discussed above, quality observation-based setting measures require quality observation systems. We sought to enhance reliability through a carefully-designed observation system with clear operational definitions, trained observers, and well-established systems of data recording.

To assess the quality of afterschool settings, observational data were collected five times across the academic year—twice in the fall with 4–6 weeks between observations, once in January–February, and twice in the spring in 4 to 6-week intervals. Approximately half of the observations in fall, winter, and spring were conducted with two observers per site. Given the variability of activities and staff in afterschool we thought that multiple measurements on varying days with varying observers would be necessary to characterize the nature of these afterschool programs accurately (Raudenbush, Martinez, Bloom, Zhu, & Lin, 2011). Observers were instructed that on each visit they should observe the full range of activities offered that day and as broad a sample of the staff and students as possible. The mean number of observers per site-visit was 1.7. On average, 1.97 staff members and 2.95 activities were rated for each combination of site, occasion, and observer. In total, 722 ratings were obtained for each CIS item and 1,092 ratings for each PPRS item, which served as the analytic sample for this study. A total of 14 observers collected the data reported here. Observers for each site varied across rounds.

⁴Two dimensions (i.e., supportive relationships with adults & appropriate structure) included more than four exemplars. For these dimensions, we selected four exemplars that best matched the purpose of our behavioral intervention study.

⁵Most frequently observed activities included homework, sports, arts or crafts, free play, and snack time. Most activities focused on improving children's interpersonal, physical/athletic, artistic, and/or problem-solving skills.

Observers participated in an initial training consisting of 2 8-hour sessions that included an overview of the constructs and measures, showing videos illustrating the concepts, and scoring practices and discussion. Training was conducted by project research staff and the university survey research center that managed the data collection effort. The training also included live-practice in afterschool programs. Two additional booster trainings occurred before the winter and spring data collection periods. In the first year, the data manager conducted a portion of the observations with the data collectors so that observer scores could be verified with a “Standard Coder.” Beginning with the second year, the data collectors were required to attend a refresher training for each wave of data collection and pass a reliability test on which they had to attain at least 80% agreement (within one scale point) with scores agreed upon by research team members, known as our Gold Standard Video (GSV) process (Stuhlman, Hamre, Downer, & Pianta, 2010). The GSV scores were established by consensus of the research team after viewing and discussing the videos and establishing agreement of at least 80% and often 90% within the research team of 3–4 members. Data collectors scoring beneath 80% on any scale received a call from the GSV coordinator, who would inquire about why certain scores were given and discuss the GSV score and the rationale. This process was completed before the deployment of observers to the field. This served as a way to prevent drift and promote the reliability and accuracy of observer ratings (Stuhlman et al., 2010). The data collectors used mini-laptop computers to record the observations in real time and then transmitted the data to the project offices.

Analysis and Results

Statistical analyses were conducted in the following steps: (1) descriptive and internal consistency reliability analyses; (2) intra-class correlation coefficients to examine inter-rater reliability; (3) confirmatory factor analyses (CFA) to evaluate the construct validity of each scale, followed by multiple group CFAs for different types of afterschool programs; (4) finally, generalizability theory analyses of reliability in light of multiple sources of measurement error. Explanations of the analytic procedures for each step are detailed below in the corresponding sections, followed by the presentation of the results.

Descriptive Data and Internal Consistency Reliability

The scales and items were evaluated in terms of descriptive statistics and internal consistency (Cronbach’s alpha). Here alpha reflects the coherence of an observer’s ratings, for a single segment (i.e., staff or activity) in a single program at a single time, across the items assessing that concept (N=722 for CIS; N=1,092 for PPRS). We used the conventional criteria of 0.70 for an acceptable level of coefficient alphas (Nunnally & Bernstein, 1994).

Caregiver Interaction Scale—Table 2 presents descriptive statistics for each of 23 CIS items and Cronbach’s alpha within each subscale, reflecting the original 4-dimension model of the CIS. The means and standard deviations of the Harshness items were smaller than those of other items in a scale, reflecting that these actions were not as common as the other behaviors. Cronbach’s alpha estimates were adequate for all CIS subscales, ranging from 0.79 to 0.90, and the entire scale, merging all subscales, had a high alpha of 0.91.

Promising Practices Rating Scale—Table 3 presents descriptive statistics for PPRS items and Cronbach's alpha within each of the hypothesized five subscales of the PPRS. All five dimensions of the PPRS achieved acceptable alpha values (0.76 to 0.89). The overall Cronbach's alpha estimate for the entire PPRS scale was 0.86., indicating good internal consistency.

Inter-Rater Reliability

The level of agreement between observers is a crucial aspect of the quality of our measures. Inter-rater reliability was assessed with intra-class correlation coefficients (ICC) reflecting the percent of variance shared across observers rating the same program at the same time, rather than unique to one observer. The reliability associated with the ICC values depends on the number of observers used in any study, with more observers yielding more reliability assessments. We therefore illustrate the implications of ICCs by reporting reliabilities computed both for a single observer and for the average number of observers in our data (1.7) (McGraw & Wong, 1996).

The results of the analyses of inter-rater reliability for the CIS, the PPRS, and their subscales appear in Table 4. The reliability for a measure that is the average of two raters was 0.77 for overall measures of both CIS and PPRS. The analyses of subscales yielded average inter-rater reliabilities (for an average of two raters) ranging from 0.59 to 0.75 for the CIS subscales and from 0.56 to 0.69 for the PPRS subscales. According to criteria proposed by Fleiss (1981) and Cicchetti and Sparrow (1981) (i.e., <0.40, poor; 0.40~0.59, fair; 0.60~0.74, good; >0.74, excellent), inter-rater reliability values for a majority of our scales and subscales were good or above. Note that the ICC values imply that measures based on observations by a single rater would have less satisfactory reliability.

Construct Validity: Confirmatory and Exploratory Factor Analysis

High internal consistency does not necessarily imply that all items on a scale are measuring a common unidimensional construct (Cortina, 1993). It is thus important to evaluate the underlying dimensional structure of our scales using factor-analytic methods in order to establish their construct validity.

We tested the consistency of our data with the hypothesized factor structures of the CIS and PPRS through confirmatory factor analysis (CFA) using AMOS 20 software (Arbuckle, 2011). We employed a bootstrapping procedure because several items in both instruments were distributed non-normally (Curran, West, & Finch, 1996). The results of the CFAs indicated little discrepancy between the bootstrap estimates and their original ML counterparts, so we report only the results of the standard ML estimation based on the normality assumption. There was a fairly high rate of missing cases for the PPRS item 11, "staff cooperation and consistency" (25.7%), because the item was "not applicable" to afterschool sites with only one staff member present at the time of observation, which we addressed through the multiple imputation procedures of AMOS, using 5 imputed datasets.

We evaluated the adequacy of fit for each model using indices commonly recommended in the literature (Schreiber, Stage, King, Nora, & Barlow, 2006). In the present study we

adopted cut-off values of $>.90$ for goodness-of-fit index (GFI) and $>.80$ for adjusted goodness-of-fit index (AGFI), and we interpreted Tucker-Lewis Index (TLI) and comparative fit index (CFI) values greater than $.90$ and $.95$ as acceptable to good fits to the data (Clara, Cox, & Enns, 2001; Long & Perkins, 2003). For standardized root mean square residual (SRMR), we considered values below 0.8 as acceptable (Hu & Bentler, 1999). For root mean-square error of approximation (RMSEA), we used criteria suggested by Browne and Cudeck (1992): $.05$ or less, good fit; $.05$ to $.08$, reasonable fit; and values greater than 0.1 , poor fit.

In case the result of the CFA did not support the a priori factor models, we conducted an exploratory factor analysis (EFA). The EFA analyses used a random split-half of the full sample to identify a factor model, the fit of which we then tested through a CFA using the other half of the sample.

In addition, we were interested in determining whether the measurement models held across different types of afterschool programs. We split the sample programs into two groups based on the socio-demographic compositions of student body, using the criterion of economic disadvantage. About 40% of the afterschool programs in the sample (17 of 44) were classified as economically disadvantaged due to more than 50% of the student population being eligible for free or reduced-price meals. These programs also served predominantly ethnic minority students (16 of 17 for the disadvantaged group versus 4 of 27 for the advantaged group) and urban populations (13 of 17 for the disadvantaged group versus 1 of 27 for the advantaged group). Thus, the contrast between these sets of programs provides an appropriate test of whether previous examinations of the reliability and validity of the PPRS generalize beyond previous assessments for programs serving disadvantaged, urban, minority populations.

We performed a series of nested multiple group CFA to test the equivalence of factor pattern, factor loadings, factor intercorrelations, and residual variance across afterschool programs serving more disadvantaged students and relatively advantaged students. We began with a baseline model with no constraints across the two sets of programs except that the same factor structure was applicable to both groups (Model 1), and we subsequently tested three increasingly restrictive models. Model 2 constrained factor loadings to be equal; Model 3 added the constraint of invariant factor variance/covariance; and Model 4 included the constraints from Model 3 plus the restriction of equal error variances across the two groups (Vandenberg & Lance, 2000). To compare the relative fit of the nested models, we relied on changes in CFI index, following Cheung and Rensvold's (2002) guideline of $.01$ or smaller as support for invariance.

Caregiver Interaction Scale—Table 2 shows the standardized factor loadings obtained from the CFA of the a priori 4-factor model from Arnett (1989) (i.e., sensitivity, harshness, detachment, and permissiveness). Items 12 and 20 had low factor loadings of 0.17 and 0.29 , respectively, which indicates that these items do not significantly contribute to their hypothesized factor. The poor performance of item 12 is understandable, in that observers almost never reported negative physical contact (mean = 1.01 on a 1 to 4 metric). In addition, the correlation between the Sensitivity and Detachment latent variables was $-.97$,

indicating a lack of discriminant validity of these two dimensions. The model also yielded less than adequate fit indices; specifically $\chi^2(224, 722)=1562.13$ ($p<.001$), GFI=0.835, AGFI=0.796, TLI=0.843, CFI=0.861, RMSEA=0.091 (90% CI: 0.087~0.095), and SRMR=0.082.

Since the hypothesized four-factor model was not supported by our data, we performed an EFA on the first half ($n=364$) of the randomly split dataset (see Appendix A, Table A1 for descriptive statistics) using principal axis factoring with promax oblique rotation. The eigenvalues for the first four factors obtained were 5.92, 2.08, 0.90, and 0.59. We retained the third factor because it proved to exceed Kaiser's criterion (i.e., eigenvalues greater than 1) after rotation. The total variance accounted for by the three factors was 49.85%. Both pattern and structure matrices (Appendix A, Table A2) showed a relatively clear three-factor solution: factor 1 consists of a combination of the original model's Sensitivity and Detachment subscales (CIS-SD); factor 2 represents the Harshness subscale (CIS-H); and factor 3 corresponds to the Permissiveness subscale (CIS-P). Most items loaded strongly on their expected factors, but there were some exceptions with weak or mixed loadings. Items 12 and 20 did not strongly load on any factors, item 6 had a substantial and inappropriate secondary loading, and item 19 loaded highest on a factor other than the intended one.

We re-ran the EFA after removing those 4 items with problematic loadings. For this EFA, three factors accounted for 53.77% of the total variance (the first factor 33.41%, the second 12.11%, and the third 8.25%). All 19 items showed unique factor loadings of more than .4 on their intended factors (Appendix A, Table A3). We then replicated this EFA solution via CFA with the second half of the sample ($n=358$). The fit results of the CFA for this 3-factor model with 19 items indicated an adequate fit to the data—specifically, $\chi^2(149, 722)=469.31$ ($p<.001$), GFI=0.878, AGFI=0.845, TLI=0.914, CFI=0.925, RMSEA=0.078 (90% CI: 0.07~0.086), and SRMR=0.061. The EFA results indicate that the remaining modest lack of fit stems from secondary cross-loadings of items on other factors, which do not appear problematic in relation to much stronger primary loadings. As displayed in Table 5, all of the standardized factor loading for the 19 items were above the conventional cut-off value of .5, and 13 items had values of .7 or higher. The refined subscales yielded very favorable Cronbach's alphas of 0.92, 0.83, and 0.87 for CIS-SD, CIS-H, and CIS-P, respectively. The correlation statistics indicated that the CIS-SD subscale had moderate correlations with the CIS-H (−0.36) and the CIS-P (−0.53), but the CIS-H and the CIS-P showed minimal intercorrelation (−0.05).

The results of multiple group CFAs indicate that the factor configuration is invariant across the two types of afterschool programs. The baseline model showed an adequate fit with CFI=.923, RMSEA=.054 (90% CI .05~.058), and SRMR=.059. The incremental changes in CFI between the two nested models were smaller than .01 for all levels of invariance, providing support for invariance in factor loadings, factor variance/covariance, and error variance across the two groups (See Appendix A, Table A4 for full results).

Promising Practices Rating Scale—Table 3 shows the standardized factor loadings and item R^2 obtained from the CFA of the hypothesized 5-factor model of the PPRS, i.e., supportive relationship with adults (PPRS-SRA), supportive relationship with peers (PPRS-

SRP), appropriate structure (PPRS-AS), level of engagement (PPRS-LE), and Chaos (PPRS-C). For all PPRS items, standardized factor loadings were above the conventional cut-off value of .5, ranging from 0.52 to 0.90. In particular, all items in the PPRS-SRP and PPRS-C subscales had factor loadings greater than 0.80. The values of key fit indices indicated that the proposed five-factor model of the PPRS scale was an adequate fit for the data in our afterschool sample; specifically, GFI=0.923, AGFI=0.892, TLI=0.925, CFI=0.94, RMSEA=0.033 (90% CI: 0.032~0.034), and SRMR=0.045. The correlations among subscales were moderate to moderately high, with the lowest correlation between PPRS-SRP and PPRS-C (-0.53) and the highest between PPRS-SRA and PPRS-AS (0.73).

Having established a 5-factor model in the overall sample, we examined measurement and structural invariance of the PPRS across afterschool programs of different composition. The results are given in Appendix A (Table A5). The baseline model yielded an adequate fit with CFI=0.923, RMSEA=0.059 (90% CI .055~.063), and SRMR=.051, suggesting that the factor pattern was consistent between afterschool programs serving more advantaged and disadvantaged populations. There was no significant change in CFI between Model 1 and Model 2, which indicates no decrement in fit and thus supports the interpretation of factor loading invariance. Comparing Model 2 and Model 3, the incremental change in CFI was also less than the cut off value of .01, which provides support for the invariant factor variance/covariance across the two groups. However, the difference in CFI between Model 3 and Model 4 was above the threshold of .01, indicating a significant reduction in model fit. Since the assumption of invariance in error variance did not hold for the PPRS, we examined modification indices and identified the errors for items 1, 2, 4, 8, 9, 12, 14, and 15 as lacking invariance. The additional analysis in which error variance for these items was not constrained to be equal yielded a CFI change of less than .01, confirming partial invariance in error variance across the two groups.

Generalizability Theory-based Reliability

Given that our afterschool program quality measures derive from a complex measurement strategy involving multiple observers and multiple observational visits, we applied generalizability (G) theory in addition to those more traditional analyses based on classical test theory. While classical test theory treats error variance as undifferentiated random variation, G theory partitions error variance into multiple sources. Doing so enables us to assess the relative contributions to the reliability of measurement of each of the different sources of error and their combinations, versus the genuine variation between programs. This information, in turn, informs decisions about how the quality and efficiency of behavioral observations can be maximized through measurement design (Brennan, 2010; Raudenbush et al., 2011; Shavelson & Webb, 1991).

In the present study, our concern is estimating qualities of after-school programs assessed through multiple observations, each of which is a joint product of the judgment of the observer and the specific events that occurred at that time. Each observer would make somewhat different judgments, and the events to be observed will differ from day to day. Thus, in research on the quality of after-school programs, we are interested in determining the impact of observers, occasions (which we refer to as rounds of assessment), and their

combinations on the reliability of our observational measures.⁶ More specifically, our measurement design allowed us to estimate five separate sources of variance: (a) between-site variance, which corresponds to true score variance in classical test theory given that site is the target of measurement; (b) observer variance, reflecting observer-to-observer differences in their average ratings of afterschool program quality; (c) variance across rounds, nested within site, which represents the changes in the ratings of the quality of any program from one round to another; (d) site-by-observer interaction, indicating the extent to which between-site differences in the ratings varied across observers; and (e) error variance confounded with the interaction between observer and round nested within site, which reflects the extent to which between-observer differences varied across rounds within sites.

We used repeated-measures analysis of variance framework to conduct a G-study of these sources of variation. The resulting variance component estimates were then used to calculate reliability coefficients under the given measurement condition and also with hypothetical variation in numbers of observers and rounds. The latter is referred to as a decision study or D-study. We calculated two types of reliability coefficients: generalizability coefficient for relative decisions and dependability index for absolute decisions (Shavelson & Webb, 1991). Appendix B provides greater detail on generalizability theory-based methods.

As mentioned earlier, each of the afterschool programs in our study was visited five times for site observations. Among the five rounds of data collection, the first two rounds were intended to establish pre-intervention characteristics of the after-school program while the last two rounds were used to obtain stable estimates of post-intervention program characteristics. To conduct a generalizability study independent from the influences of the intervention, we stacked pre- and post-intervention data after assigning new site IDs to post-intervention data, thereby treating the pre- and post-data as though they represented different sites. The data from the interim round (3rd round) were not included in the current analysis. As a result, the analysis includes a total of 276 data points (44×2 Sites, 2 Rounds, 1~2 Observers).

Results of the G-study—Table 6 presents the variance component estimates for each of three CSI and five PPRS scales as well as the reliability coefficients obtained for our research design. The amount of variance attributable to each component varied across scales. The “true score” variance of interest, site (σ_s^2), accounted for 11% to 43% of total variance across the CIS and PPRS scales. Variance across observers’ mean ratings (σ_o^2) was smaller than the other components, accounting for 1% to 18% of variance. In contrast, a relatively large percentage of variance was attributable to day to day variation in program functioning, as reflected in round nested within site ($\sigma_{r:s}^2$), ranging from 14% to 39%. The site-by-observer interaction term (σ_{so}^2) contributed almost negligible variance to the model for some scales (0% ~ 3.7%), but explained a modest proportion of variance for others (12%

⁶Items within measures could be treated as an additional source of error, as well. Now that we have established adequate internal consistency and dimensional structure, however, the item-level error will be constant. Therefore, it will be sufficient to ignore that source and allow it to be incorporated in the undifferentiated residual error variance.

~ 18%). The last component, observer-by-round interaction plus error ($\sigma_{o(r:s),e}^2$), explained 23% to 38% of total variance.

Table 6 also reports the two types of reliability coefficients calculated given the number of rounds and observers for the current research design. Reliability coefficients varied across the scales, with generalizability coefficients ranging from 0.26 to 0.69 and dependability indices ranging from 0.25 to 0.67. There was little difference between generalizability coefficients and dependability indices because the variance attributable to observer was small relative to the other components. Detailed variance component results specific to each scale are presented below.

For CIS-SD, CIS-H, and PPRS-SRA, the largest proportion of observed score variance was accounted for by site (35~43%), indicating that these afterschool programs differed systematically in terms of staff's sensitivity-detachment and harshness and overall relationship between staff and youth. The second largest proportion of variance was attributed to the observer-by-round interaction plus unexplained variance (23%~32%). The next largest proportion of variance was explained by round nested within site (20%~27%), indicating that there were fairly substantial changes in the ratings of program quality from round to round within sites. The variances due to observer and site-by-observer interaction were relatively small, revealing that the relative standing of afterschool programs was constant across observers. The small proportion of variance attributable to observer (4%~10%) also indicates that there were minimal differences in the mean ratings of these three scales across observers. Reliability coefficients were relatively high for these scales. Under the given measurement condition, generalizability coefficients were 0.64, 0.68, and 0.69, and dependability indices were 0.62, 0.63, and 0.67, for PPRS-SRA, CIS-SD, and CIS-H, respectively.

For PPRS-C and PPRS-AS, a substantial amount of variance was explained by site (32% and 24%), but the largest amount of variance was attributable to the observer-by-round interaction confounded with residual error (35% and 38%). Another large source of error was the variance due to round-to-round differences within a site (21% and 25%). The site-by-observer variance was negligible for PPRS-C (0.8%), but fairly substantial for PPRS-AS (15%). The proportion of variance attributable to observer was small for both scales, 1% for PPRS-AS and 5% for PPRS-C. Given the conditions of our measurement procedure, both generalizability coefficients and dependability indices were about 0.6 for PPRS-C and around 0.5 for PPRS-AS.

The last three scales, CIS-P, PPRS-SRP, and PPRS-LE, consisted of a relatively small percentage of variance attributable to sites (13%, 17%, and 11%, respectively) and thus a relatively larger percentage of variance due to the error components. This led to low reliability coefficients, with generalizability coefficients of 0.26 (PPRS-LE), 0.39 (CIS-P), and 0.42 (PPRS-SRP), and dependability indices of 0.25 (PPRS-LE), 0.33 (CIS-P), and 0.38 (PPRS-SRP). In particular, for PPRS-SRP and PPRS-LE, the between-round variance and observer-by-round interaction plus error together explained roughly two-thirds of the observed score variance. For CIS-P, the between-observer variance and site-by-observer variance together accounted for 35% of the total variance, much more than the other scales.

Results of the D-study—Using the variance component estimates obtained from the generalizability study, we calculated dependability indices for different numbers of observers (1 to 10) and different numbers of rounds within a site (1 to 4). Figure 1 shows how dependability indices will change across different combinations of numbers of observers and rounds for each of eight observation scales. It is evident that a single round of observation would not produce an adequate level of reliability. Figure 1 also shows that increasing the number of observers and/or the number of rounds raises the dependability index, but that returns diminish as the number of either increases beyond a few. For example, the estimated dependability index of the CIS-H scores is about 0.80 for a four-observer and four-round design and around 0.85 for a ten-observer and four-round design.

However, the extent of gains achieved by adding the number of rounds or observers varies somewhat across scales, reflecting which error variance components are largest for each of them. Particularly notable is that the CIS-P scale benefits more by increasing the number of observers than the number of rounds while the opposite is generally the case for other scales. Figure 2 illustrates that the addition of observers and rounds has substantially different impacts on CIS-P and PPRS-LE, for example.

Discussion

Assessment of settings is a more complex enterprise than assessment at the individual level. Given the importance of social ecological settings for youth development, however, it is as important to assess the quality of these settings as units in their own right as it is to measure individual-level features (Raudenbush & Sampson, 1999). As today's youth spend a considerable amount of time in afterschool settings, these settings are increasingly viewed as a major social context for youth development, and they are of great interest to parents, educators, researchers and policymakers. The quality of afterschool settings is particularly relevant to school practices and policies, given a wealth of research evidence that teaching and learning processes could be enhanced or impeded by what students bring into class (Natriello & Dornbusch, 1983; Warren et al., 2006). Students' behaviors, attitudes, skills and competencies are likely improved by positive experiences and activities available in high quality afterschool settings, which would, in turn, lead to creating better contexts for teaching and learning in school.

In response to a growing need for tools to assess the quality of afterschool settings (Granger, 2005), the present study examined whether the CIS and PPRS could serve as reliable and valid measures of the quality of afterschool settings, and the results are promising. Since its publication (Arnett, 1989), the CIS has been widely used to rate caregiver-child interactions in early child care research (Austin, Lindauer, Rodriguez, Norton, & Nelson, 1997; Campbell & Milbourne, 2005; de Kruif, McWilliam, Ridley, & Wakely, 2000). Most of this research has reported good interrater reliability and internal consistency, but despite its wide use, its hypothesized factor structure or dimensionality has been relatively unexplored.

Our analyses revealed that a three-factor solution best described our data, indicating that the dimensions underlying the CIS are sensitivity-detachment, harshness, and permissiveness, merging the sensitivity and detachment factors from the original factor structure. According

to Arnett (1989), the sensitivity subscale reflects warmth, enthusiasm, and developmentally appropriate interactions while the detachment factor reflects uninvolvement, indifference, and lack of interest and engagement. Our result suggests that sensitivity and detachment represent the opposite ends of the same continuum rather than two distinct dimensions, consistent with an earlier finding of an unduly high negative correlation between the two (de Kruif et al., 2000). Also consistent with previous research, our results demonstrated good internal consistency and satisfactory interrater reliability. In combination, these results suggest that this observational scale of caregiver-child interactions within early child care settings can be reliably and validly applied to afterschool settings that serve older children and youth.

While the CIS focuses on staff-child interactions, the PPRS measures broader aspects of afterschool program processes and practices, including staff-child relationships, peer relationships, level of engagement, and appropriate structure and supervision. Afterschool research has recently begun to adopt the PPRS (Mahoney et al., 2007; Vandell, O'Cadiz, & Hall, 2009), but little is known about its psychometric properties. Our analyses revealed that the hypothesized five-factor model fit the modified version of the PPRS very well, and each subscale displayed good internal consistency. Like the CIS, the PPRS also showed satisfactory performance in terms of interrater reliability, which suggests that our observation system was successful in training the observers to use the scales in a similar manner.

Well-validated setting-level measurement tools can be used for both research and practice purposes. For research, they can be used to enhance our understanding of how afterschool programs work and how to make them work better (Granger, 2005). For practice, validated measurement tools can be employed to provide program directors and staff information about areas of strengths and weaknesses in their practice and to give them practical guidance on what changes should be made to better serve their students. This will help hold afterschool programs accountable for ongoing assessment of their practice and continuous improvement (Granger, Durlak, Yohalem, & Reisner, 2007). Furthermore, we found that the factor structure differed little between programs serving primarily disadvantaged, urban, and minority populations and those serving mainly advantaged non-minority populations in suburban and rural settings. This result suggests that these instruments may have relatively broad applicability for after-school settings.

The findings of this study confirm the potential promise of the instruments, on the one hand, and reveal a critical new issue for improving measurement design, on the other hand. Observation data collection involves multiple sources of measurement error, and our research has revealed one that has received little attention in prior work in this area. Day-to-day fluctuations in the events observed were often bigger sources of variation in scores than were differences between observers and between items comprising the measures. Although strategies to reduce between-observer inconsistencies may be needed for CIS-P, this was not true for other measures. In general, an increase in the number of rounds of observation is of greater utility rather than an increase in the number of observers. Further, additional rounds bring definite improvement to the reliability of virtually all of the subscales simultaneously assessed by each measure, where additional observers benefit a smaller portion of subscales.

Thus, the results of the generalizability and decision analyses suggest that future assessment designs should place more weight on increasing the number of occasions that afterschool programs are observed, rather than the number of observers. Increasing the number of segments per round of observation may be an equally valid way of improving reliability. Indeed, the PPRS observational protocol recommends at least six segments of observation per visit (Vandell et al., 2004), while in our observation system 3 successive activities were observed for each round.

The current study has several limitations that should be addressed in future research concerning the measurement of afterschool program quality. First, although direct observation has obvious merits as a method for measuring setting-level quality, it is restricted to the observable aspects of quality, i.e., overt behaviors and concrete events. Psychological and perceptual processes such as social norms, sense of belonging, emotional engagement are not readily captured using observational approaches. Rather, survey-based methods are more suitable for gathering such information. Thus, more complete assessment of settings should include both direct observation and survey methods (Raudenbush & Sampson, 1999). It would be desirable if observational measures were further validated using survey-based measures. Second, our study did not include three subscales of the PPRS, so additional research is needed to extend our work to those portions of this measure. Given that an ultimate goal of afterschool interventions is to improve youth outcomes through the enhancement of the quality of afterschool settings, our observational measures definitely need to be further validated by evidence of associations with a range of important youth and staff outcomes. Similarly, future research also needs to explore the extent to which the measures of afterschool program quality mediate the effects of interventions.

We believe that our research adds much needed evidence on the measurement properties of observational instruments for after-school settings, particularly given the importance of program quality for youth behavioral and academic outcomes. Our research addresses two widely known measures of dimensions of program quality prominent in research on program effectiveness, and our sample includes programs serving diverse populations in terms of race-ethnicity, urban to rural locales, and socio-economic status. Findings from generalizability analyses indicate future directions for improving observational measurement that help inform how to obtain more accurate and informative assessment of afterschool program quality.

Acknowledgments

The researchers would like to acknowledge funding support from the William T. Grant Foundation [Grant # 8529]; the Wallace Foundation [grant number 20080489]; and the National Institute for Drug Abuse [Grant # R01 DA025187]. We also gratefully acknowledge the involvement of Lisa White and the Pennsylvania State University Survey Research Center; the team of data collectors; the participating afterschool staff, students, and their parents; and Daniel Perkins, Jason Rose, Jill Schulte, Lindsay Unger, Sharon Childs, Linda Halgunseth, and Alba Blandino who assisted in the coordination and collection of these data.

References

- Arbuckle, J.L. *IMB SPSS Amos 20 user's guide*. Armonk, NY: IBM; 2011.
- Arnett J. Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*. 1989; 10:541–552.

- Austin AMB, Lindauer SLK, Rodriguez A, Norton ML, Nelson FAG. Conditions of caregiving, provider nurturance and quality care. *Early Child Development and Care*. 1997; 135(1):21–33. [10.1080/0300443971350103](https://doi.org/10.1080/0300443971350103)
- Blazevski, J.; Smith, C. After-school quality and school-day outcomes in Michigan's 21st CCLC program. High/Scope Educational Research Foundation; Ypsilanti, MI: 2007.
- Brennan RL. Generalizability theory and classical test theory. *Applied Measurement in Education*. 2010; 24(1):1–21.
- Bronfenbrenner, U. The ecology of human development: experiments by nature and design. Cambridge, Mass: Harvard University Press; 1979.
- Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociological Methods & Research*. 1992; 21(2):230–258.
- Campbell PH, Milbourne SA. Improving the quality of infant-toddler care through professional development. *Topics in Early Childhood Special Education*. 2005; 25(1):3–14.
- Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*. 2002; 9(2):233–255.
- Cicchetti D, Sparrow SS. Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency*. 1981; 86:127–137. [PubMed: 7315877]
- Clara IP, Cox BJ, Enns MW. Confirmatory factor analysis of the depression-anxiety-stress scales in depressed and anxious patients. *Journal of Psychopathology and Behavioral Assessment*. 2001; 23(1):61–67.
- Cortina JM. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*. 1993; 78(1):98–104.
- Cronbach LJ. Coefficient Alpha and the internal structure of tests. *Psychometrika*. 1951; 16(3):297–334.
- Cronbach, LJ.; Gleser, GC.; Nanda, H.; Rajaratnam, N. The dependability of behavioral measurement. New York: Wiley; 1972.
- Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*. 1996; 1(1):16–29.
- de Kruif REL, McWilliam RA, Ridley SM, Wakely MB. Classification of teachers' interaction behaviors in early childhood classrooms. *Early Childhood Research Quarterly*. 2000; 15(2):247–268. [http://dx.doi.org/10.1016/S0885-2006\(00\)00051-X](http://dx.doi.org/10.1016/S0885-2006(00)00051-X).
- DuBois D, Holloway B, Valentine J, Cooper H. Effectiveness of mentoring programs for youth: A meta-analytic review. *American Journal of Community Psychology*. 2002; 30(2):157–197. [10.1023/a:1014628810714](https://doi.org/10.1023/a:1014628810714) [PubMed: 12002242]
- Durlak JA, Weissberg RP, Pachan M. A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology*. 2010; 45(3):294–309. [10.1007/s10464-010-9300-6](https://doi.org/10.1007/s10464-010-9300-6) [PubMed: 20300825]
- Eccles, JS.; Gootman, JA. Community programs to promote youth development. Washington, D. C: National Academy Press; 2002.
- Eccles, JS.; Roeser, RW. Schools, academic motivation, and stage-environment fit. In: Lerner, RM.; Steinberg, L., editors. *Handbook of adolescent psychology*. 3. Hoboken: Wiley; 2009. p. 404–434.
- Fleiss, JL. The measurement of interrater agreement. In: Fleiss, JL., editor. *Statistical methods for rates and proportions*. New York: Wiley & Sons; 1981. p. 212–236.
- Franco N, Levitt MJ. The social ecology of middle childhood: Family support, friendship quality, and self-esteem. *Family Relations*. 1998; 47(4):315–321.
- Fraser BJ. Classroom environment instruments: Development, validity and applications. *Learning Environments Research*. 1998; 1:7–33.
- Gerstenblith, SA.; Soule, DA.; Gottfredson, DC.; Lu, S.; Kellstrom, MA.; Womer, SC.; Bryner, SL. After-school programs, antisocial behavior, and positive youth development: An exploration of the relationship between program implementation and changes in youth behavior. In: Mahoney, J.; Larson, R.; Eccles, J., editors. *Organized activities as contexts of development: Extracurricular activities, after-school and community programs*. Mahwah, NJ: Lawrence Erlbaum Associates; 2005. p. 457–477.

- Gonzales NA, Cauce AM, Friedman RJ, Mason CA. Family, peer, and neighborhood influence on academic achievement among African American adolescents: One year prospective effects. *American Journal of Community Psychology*. 1996; 24:365–387. [PubMed: 8864209]
- Gorman-Smith D, Tolan PH, Henry DB. A developmental-ecological model of the relation of family functioning to patterns of delinquency. *Journal of Quantitative Criminology*. 2000; 16(2):169–198.
- Gortmaker SL, Lee RM, Mozaffarian RS, Sobol AM, Nelson TF, Roth BA, Wiecha JL. Effect of an after-school intervention on increases in children's physical activity. *Medicine & Science in Sports & Exercise*. 2012; 44(3):450–457. [PubMed: 21814151]
- Gottfredson D, Gerstenblith S, Soulé D, Womer S, Lu S. Do after school programs reduce delinquency? *Prevention science*. 2004; 5(4):253–266.10.1023/b:prev.0000045359.41696.02 [PubMed: 15566051]
- Granger, RC. Improving the settings that affect youth. William T. Grant Foundation; New York: 2005. Retrieved from http://www.wtgrantfoundation.org/about_us/annual_reports
- Granger, RC.; Durlak, JA.; Yohalem, N.; Reisner, E. Improving after-school program quality. William T. Grant Foundation; New York: 2007.
- Grossman JB, Goldsmith J, Sheldon J, Arbretton AJA. Assessing after-school settings. *New Directions for Youth Development*. 2009; 121:89–108.10.1002/yd.298 [PubMed: 19358187]
- Halle, T.; Vick Whittaker, JE.; Anderson, R. Quality in early childhood care and education settings: A compendium of measures. Child Trends; Washington, DC: 2010.
- Hamre, BK.; Pianta, RC.; Mashburn, AJ.; Downer, JT. Building a science classrooms: Application of CLASS framework in over 4, 000 U.S. early childhood and elementary classrooms. 2007. Retrieved from <http://www.icpsr.umich.edu/files/PREK3RD/resources/pdf/BuildingAScienceOfClassroomsPiantaHamre.pdf>
- Harms, T.; Jacobs, E.; White, D. School age care environment rating scale. Teachers College Press; New York, NY: 1996. Retrieved from <http://search.proquest.com/docview/62666810?accountid=13158>
- Hu LT, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999; 6:1–55.
- Kirtland KA, Porter DE, Addy CL, Neet MJ, Williams JE, Sharpe PA, Ainsworth BE. Environmental measures of physical activity supports: Perception versus reality. *American journal of preventive medicine*. 2003; 24(4):323–331. [http://dx.doi.org/10.1016/S0749-3797\(03\)00021-7](http://dx.doi.org/10.1016/S0749-3797(03)00021-7). [PubMed: 12726870]
- Kuperminc GP, Smith EP, Henrich CC. Introduction to the special issue on social and motivational processes in after-school settings: Bridging gaps between theory, research, and practice. *The Journal of early adolescence*. 2013; 33(1):5–16.10.1177/0272431612469055
- Larson R. Toward a psychology of positive youth development. *American Psychologist*. 2000; 55:170–183. [PubMed: 11392861]
- Lauer PA, Akiba M, Wilkerson SB, Aporthorp HS, Snow D, Martin-Glenn ML. Out-of-school-time programs: A meta-analysis of effects for at-risk students. *Review of Educational Research*. 2006; 76(2):275–313.10.3102/00346543076002275
- Long DA, Perkins DD. Confirmatory factor analysis of the sense of community index and development of a brief SCI. *Journal of Community Psychology*. 2003; 31(3):279–296.
- Mahoney JL, Parente ME, Lord H. After-school program engagement: Links to child competence and program quality and content. *The Elementary School Journal*. 2007; 107(4):385–404.10.1086/516670
- Mahoney JL, Stattin H. Leisure activities and adolescent antisocial behavior: The role of structure and social context. *Journal of Adolescence*. 2000; 23(2):113–127. <http://dx.doi.org/10.1006/jado.2000.0302>. [PubMed: 10831137]
- Mahoney JL, Stattin H, Lord H. Unstructured youth recreation centre participation and antisocial peers. *International Journal of Behavioral Development*. 2004; 28:553–560.
- Mahoney JL, Zigler EF. Translating science to policy under the No Child Left Behind Act of 2001: Lessons from the national evaluation of the 21st-Century Community Learning Centers. *Journal of Applied Developmental Psychology*. 2006; 27(4):282–294. <http://dx.doi.org/10.1016/j.appdev.2006.04.001>.

- Maimon D, Browning CR. Unstructured socializing, collective efficacy, and violent behavior among urban youth. *Criminology*. 2010; 48(2):443–474.10.1111/j.1745-9125.2010.00192.x
- McCall GJ. Systematic field observation. *Annual Review of Sociology*. 1984; 10:263–282.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*. 1996; 1(1):30–46.
- McLaughlin, MW.; Irby, MA.; Langman, J. Urban sanctuaries: neighborhood organizations in the lives and futures of inner-city youth. San Francisco: Jossey-Bass Publishers; 1994.
- Natriello G, Dornbusch SM. Bringing behavior back in: The effects of student characteristics and behavior on the classroom behavior of teachers. *American Educational Research Journal*. 1983; 20(1):29–43.10.2307/1162673
- Nock MK, Kurtz SMS. Direct behavioral observation in school settings: Bringing science to practice. *Cognitive and Behavioral Practice*. 2005; 12:359–370.
- Nunnally, J.; Bernstein, I. *Psychometric theory*. New York: McGraw-Hill; 1994.
- Osgood DW, Anderson AL. Unstructured socializing and rates of delinquency. *Criminology*. 2004; 42(3):519–550.10.1111/j.1745-9125.2004.tb00528.x
- Perkins, DF.; Borden, LM. Key elements of community youth development programs. In: Villarruel, FA.; Perkins, DF.; Borden, LM.; Keith, JG., editors. *Community youth development: Practice, policy, and research*. Thousand Oaks, CA: Sage; 2003. p. 327-340.
- Pianta, RC. Standardized observation and professional development: A focus on individualized implementation and practices. In: Zaslow, M.; Martinez-Beck, I., editors. *Critical issues in early childhood professional development*. Baltimore, MD: Brookes Publishing Company; 2005. p. 231-254.
- Pianta, RC.; Allen, JP. Building capacity for positive youth development in secondary school classrooms: Changing teachers' interactions with students. In: Shinn, M.; Yoshikawa, H., editors. *Toward positive youth development: Transforming schools and community programs*. New York: Oxford University Press; 2009. p. 21-39.
- Pianta RC, Hamre BK. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*. 2009; 38:109–119.
- Pierce KM, Bolt D, Vandell DB. Specific features of after-school program quality: Associations with children's functioning in middle childhood. *American Journal of Community Psychology*. 2010; 45:381–393. [PubMed: 20336364]
- Pierce KM, Hamm JV, Vandell DL. Experiences in after-school programs and children's adjustment in first-grade classrooms. *Child Development*. 1999; 70:56–767.
- Raudenbush, SW.; Martinez, A.; Bloom, H.; Zhu, P.; Lin, F. Studying the reliability of group-level measures with implications for statistical power: A six-step paradigm. William T. Grant Foundation; 2011. Retrieved from <http://www.wtgrantfoundation.org/resources/measures-of-social-settings>
- Raudenbush SW, Sampson RJ. Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*. 1999; 29:1–41.
- Riggs NR. After-school program attendance and the social development of rural Latino children of immigrant families. *Journal of Community Psychology*. 2006; 34:75–87.
- Schreiber JB, Stage FK, King J, Nora A, Barlow EA. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*. 2006; 99(6): 323–337.
- Shavelson, RJ.; Webb, NM. *Generalizability theory: A primer*. Newbury Park, CA: Sage; 1991.
- Shermoff DJ. Engagement in after-school programs as a predictor of social competence and academic performance. *American Journal of Community Psychology*. 2010; 45(3–4):325–337.10.1007/s10464-010-9314-0 [PubMed: 20232243]
- Shin R, Daly B, Vera E. The relationships of peer norms, ethnic identity, and peer support to school engagement in urban youth. *Professional School Counseling*. 2007; 10(4):379–388.
- Shinn, M.; Rapkin, BD. Cross-level research without cross-ups in community psychology. In: Rappaport, J.; Seidman, E., editors. *Handbook of community psychology*. New York, NY: Springer; 2000. p. 669-695.

- Skinner, CH.; Dittmer, KI.; Howell, LA. Direct observation in school settings: Theoretical issues. In: Shapiro, ES.; Kratochwill, TR., editors. Behavioral assessment in schools: theory, research, and clinical foundations. 2. New York: Guilford Press; 2000. p. 19-45.
- Smith, C.; Hohmann, C. Full findings from the youth program quality assessment validation study. High/Scope Educational Research Foundation; Ypsilanti, MI: 2005.
- Smith EP, Faulk M, Sizer MA. Exploring the meso-system: The roles of community, family, and peers in adolescent delinquency and positive youth development. Youth & society. in press.
- Stuhlman, MW.; Hamre, BK.; Downer, JT.; Pianta, RC. A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems. University of Virginia, Center for Advanced Study of Teaching and Learning; 2010.
- Vandell, DL.; O'Cadiz, P.; Hall, V. High quality supplemental educational services and afterschool partnerships demonstration project: An evaluation study of THINK Together programs in Santa Ana Unified school district. University of California; Irvine: 2009.
- Vandell, DL.; Reisner, ER.; Brown, BB.; Pierce, KM.; Dadisman, K.; Pechman, EM. The study of promising after-school programs: Descriptive report of the promising programs. 2004. Retrieved from <http://www.wcer.wisc.edu/childcare/statements.html>
- Vandell, DL.; Reisner, ER.; Pierce, KM. Outcomes linked to high-quality afterschool programs: Longitudinal findings from the study of promising afterschool programs. Policy Studies Associates; Washington, DC: 2007.
- Véronneau MH, Dishion TJ. Middle school friendships and academic achievement in early adolescence: A longitudinal analysis. The Journal of early adolescence. 2011; 31(1):99–124.10.1177/0272431610384485 [PubMed: 21552353]
- Walker, KE.; Arbreton, AJA. After-school pursuits: An examination of outcomes in the San Francisco beacon initiative. Public/Private Ventures; New York, NY: 2004. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED503196>
- Warren J, Bohanon-Edmonson H, Turnbull A, Sailor W, Wickham D, Griggs P, Beech S. School-wide positive behavior support: Addressing behavior problems that impede student learning. Educational Psychology Review. 2006; 18(2):187–198.10.1007/s10648-006-9008-1
- Yohalem N, Wilson-Ahlstrom A. Inside the black box: Assessing and improving quality in youth programs. American Journal of Community Psychology. 2010; 45(3–4):350–357.10.1007/s10464-010-9311-3 [PubMed: 20300822]

Appendix A: Supplementary Tables

Table A1

Descriptive Statistics of Random Split Sample for CIS

		EFA Sample (N=364)		CFA Sample (N=358)	
		Mean	SD	Mean	SD
Sensitivity					
1	Speaks warmly to children	3.12	0.84	2.95	0.88
2	Listens attentively to children	3.09	0.83	2.95	0.86
3	Excited about teaching	2.84	0.9	2.68	0.88
4	Encourages to try different experiences	1.85	0.87	1.87	0.94
5	Enthusiastic about children's activities/efforts	2.8	0.93	2.65	0.92
6	Encourages to exhibit prosocial behavior	2.37	0.95	2.27	0.99
7	Positive physical contact	1.85	0.86	1.82	0.88
8	Sincere in tone of voice and manner	3.34	0.75	3.15	0.85
Harshness					

		EFA Sample (N=364)		CFA Sample (N=358)	
		Mean	SD	Mean	SD
9	Seems critical of children	1.13	0.4	1.19	0.51
10	Places high value on obedience-directive	1.65	0.82	1.7	0.91
11	Speaks with irritation or hostility to children	1.24	0.52	1.27	0.57
12	Negative physical contact	1.01	0.09	1.02	0.13
13	Punishes w/o explanation or other redirection	1.08	0.32	1.11	0.39
14	Finds fault easily with children	1.1	0.39	1.14	0.45
15	Prohibit many things children want to do	1.27	0.48	1.34	0.6
Detachment					
16	Emotionally distant from children	1.7	0.87	1.78	0.87
17	Considerable time not involving children	1.77	0.84	1.89	0.87
18	Seems interested in children's activities*	1.99	0.86	2.13	0.92
19	Supervises children closely*	2.03	0.89	2.15	0.92
Permissiveness					
20	Exercises little control over children	1.52	0.69	1.58	0.74
21	Reprimands children when they misbehave*	2.3	1.06	2.4	1.04
22	Exercises firmness when necessary*	2.29	1.01	2.34	1.01
23	Expects children to exercise self-control*	2.08	0.89	2.09	0.93

Note: Items indicated by * are reverse-coded items.

Table A2

Pattern and Structure Matrix for Principal Axis Factoring Analysis of 23 CIS Items (N=364)

Item	Pattern Matrix			Structure Matrix		
	1	2	3	1	2	3
<i>Sensitivity</i>						
Speaks warmly to children	0.84	-0.11	0.10	0.84	-0.49	-0.35
Listens attentively to children	0.80	-0.03	0.02	0.81	-0.41	-0.41
Excited about teaching	0.89	0.02	0.09	0.84	-0.39	-0.38
Encourages to try different experiences	0.63	0.22	0.00	0.53	-0.08	-0.32
Enthusiastic about children's activities	0.88	-0.03	0.04	0.87	-0.44	-0.43
Encourages to exhibit prosocial beh.	0.57	0.22	-0.38	0.64	-0.07	-0.62
Positive physical contact	0.66	0.27	0.11	0.47	-0.03	-0.23
Sincere in tone of voice and manner	0.51	-0.16	-0.29	0.73	-0.41	-0.57
<i>Harshness</i>						
Seems critical of children	0.07	0.62	0.01	-0.23	0.59	0.01
Places high value on obedience-directive	0.04	0.57	-0.17	-0.13	0.54	-0.16
Speaks with irritation or hostility	-0.11	0.60	-0.15	-0.31	0.65	-0.05
Negative physical contact	-0.16	-0.02	-0.14	-0.08	0.05	-0.05
Punishes w/o explanation	0.03	0.55	-0.01	-0.22	0.53	0.01

Item	Pattern Matrix			Structure Matrix		
	1	2	3	1	2	3
Finds fault easily with children	-0.03	0.58	-0.01	-0.30	0.59	0.05
Prohibit many things child want to do	0.21	0.77	0.10	-0.20	0.68	0.04
<i>Detachment</i>						
Emotionally distant from children	-0.67	0.16	0.03	-0.76	0.47	0.40
Considerable time not involving children	-0.44	0.14	0.14	-0.58	0.35	0.38
Seems interested in children's activities*	-0.75	0.13	0.03	-0.83	0.48	0.44
Supervises children closely*	-0.18	0.09	0.66	-0.57	0.21	0.76
<i>Permissiveness</i>						
Exercises little control over children	-0.08	0.31	0.25	-0.36	0.36	0.31
Reprimands children when misbehave*	0.06	-0.07	0.84	-0.35	-0.04	0.80
Exercises firmness when necessary*	0.17	-0.01	1.00	-0.35	-0.03	0.91
Expects children to exercise self-control*	-0.09	-0.09	0.78	-0.46	-0.01	0.82

Note1: Principal axis factoring with the Promax rotation (oblique) was used because the factors were expected to be correlated with each other. The eigen-value greater-than-1-rule was used as the criterion for retaining factors.

Note2: Items indicated by * are reverse-coded items.

Note3: The bolded coefficients highlight the highest coefficients and the bolded italicized coefficients highlight those showing less than 2.0 difference

Table A3

Pattern and Structure Matrix for Principal Axis Factoring Analysis of 19 CIS Items (N=364)

Item	Pattern Matrix			Structure Matrix		
	1	2	3	1	2	3
<i>Sensitivity</i>						
Speaks warmly to children	0.86	-0.05	0.09	0.85	-0.42	-0.28
Listens attentively to children	0.82	0.00	0.01	0.81	-0.34	-0.33
Excited about teaching	0.89	0.05	0.07	0.84	-0.33	-0.31
Encourages to try different experiences	0.56	0.16	-0.03	0.51	-0.08	-0.28
Enthusiastic about children's activities	0.87	-0.02	0.02	0.87	-0.39	-0.35
Positive physical contact	0.57	0.21	0.06	0.46	-0.04	-0.19
Sincere in tone of voice and manner	0.58	-0.10	-0.25	0.73	-0.34	-0.49
<i>Harshness</i>						
Seems critical of children	0.05	0.64	0.02	-0.23	0.62	-0.03
Places high value on obedience-directive	0.05	0.59	-0.14	-0.15	0.58	-0.19
Speaks with irritation/hostility to children	-0.11	0.63	-0.12	-0.32	0.68	-0.11
Punishes w/o explanation/redirection	0.00	0.55	0.00	-0.23	0.55	-0.03
Finds fault easily with children	-0.04	0.61	0.01	-0.30	0.63	0.00
Prohibit many things child want to do	0.15	0.77	0.12	-0.22	0.70	0.01
<i>Detachment</i>						
Emotionally distant from children	-0.71	0.10	0.03	-0.77	0.40	0.33

Item	Pattern Matrix			Structure Matrix		
	1	2	3	1	2	3
Considerable time not involving children	-0.49	0.09	0.11	-0.58	0.29	0.32
Seems interested in children's activities*	-0.79	0.09	0.03	-0.84	0.42	0.36
<i>Permissiveness</i>						
Reprimands children when misbehave*	0.01	-0.01	0.85	-0.34	-0.06	0.84
Exercises firmness when necessary*	0.09	0.04	0.96	-0.34	-0.05	0.93
Expects children to exercise self-control*	-0.20	-0.11	0.67	-0.44	-0.06	0.76

Note1: Principal axis factoring with the promax rotation (oblique) was used because the factors were expected to be correlated with each other. The eigen-value greater-than-1-rule was used as the criterion for retaining factors.

Note2: Items indicated by * are reverse-coded items.

Note3: The bolded coefficients highlight the highest coefficients

Table A4

Invariance of CIS across Programs of Advantaged and Disadvantaged Populations

Model	χ^2	df	χ^2/df	CFI	RMSEA (90% CI)	SRMR	CFI
Model 1 (No constraints)	931.27	298	3.125	.923	.054 (.050~.058)	.059	--
Model 2 (Equal factor loadings)	989.85	314	3.152	.918	.055 (.051~.059)	.064	.005
Model 4 (Equal factor var/cov)	1037.97	320	3.244	.913	.056 (.052~.060)	.073	.005
Model 5 (Equal error variance)	1079.63	339	3.185	.910	0.055 (.051~.059)	.075	.003

Table A5

Invariance of PPRS across Programs of Advantaged and Disadvantaged Populations

Model	χ^2	df	χ^2/df	CFI	RMSEA (90% CI)	SRMR	CFI
Model 1 (No constraints)	1044.38	218	4.791	.923	.059 (.055~.063)	.051	--
Model 2 (Equal factor loadings)	1078.71	230	4.690	.921	.058 (.055~.062)	.054	.002
Model 3 (Equal factor var/cov)	1190.17	245	4.858	.912	.059 (.056~.063)	.060	.009
Model 4 (Equal error variance)	1361.89	262	5.198	.897	.062 (.059~.065)	.066	.015

Appendix B: Details on Generalizability Theory-based Methods

In classical test theory, reliability is defined as the proportion of observed-score variance attributable to variance in true scores, with the remainder of the variance attributable to a single source of measurement error (Cronbach, 1951). Generalizability (G) theory, introduced by Cronbach and his colleagues (1972), extended classical test theory to account for multiple sources of error. We used a repeated-measures analysis of variance framework to estimate the variance components associated with each facet of measurement and their interactions.

If our measurement design had been fully crossed, where two facets (i.e., observers and rounds) are crossed with the target of measurement (i.e., afterschool sites) and with one another, the total variation in observed scores would be expressed as in equation (1) below.

In this equation, σ_s^2 is the true score variance and the following 6 components are sources of measurement error: observer (σ_o^2), round (σ_r^2), site-by-observer interaction (σ_{so}^2), site-by-round interaction (σ_{sr}^2), observer-by-round interaction (σ_{or}^2), and the three-way interaction confounded with unmeasured error ($\sigma_{sor,e}^2$). Because rounds were nested within sites in our measurement design, the model for the variance of observed scores reduces to equation (2), with the round effect and site-by-round interaction ($\sigma_r^2, \sigma_{sr}^2$) confounded in one component ($\sigma_{r:s}^2$). Another two components ($\sigma_{or}^2, \sigma_{sor,e}^2$) were also confounded in one term ($\sigma_{o(r:s),e}^2$). Thus, our analysis of variance yielded five variance components, one representing true score variance and the other four components constituting error variance.

$$\sigma_{X_{sor}}^2 = \sigma_s^2 + \sigma_o^2 + \sigma_r^2 + \sigma_{so}^2 + \sigma_{sr}^2 + \sigma_{or}^2 + \sigma_{sor,e}^2 \quad (1)$$

$$\sigma_{X_{or(s)}}^2 = \sigma_s^2 + \sigma_o^2 + \sigma_{r:s}^2 + \sigma_{so}^2 + \sigma_{o(r:s),e}^2 \quad (2)$$

We used the VARCOMP procedure in SPSS to obtain variance component estimates. The variance component estimates were then used to calculate two types of generalizability theory-based reliability coefficients: generalizability coefficient (G-coefficient) and dependability index (D-index). A generalizability coefficient is used in contexts where the relative standing of targets, in our case, afterschool sites, is of concern and computed based on relative error variance. In contrast, a dependability index is used in situations where the absolute performance level (e.g., prespecified qualifying score) is of interest and calculated based on absolute error variance (Shavelson & Webb, 1991). While the calculation of absolute error variance involves all variance components except the one for the target of measurement, relative error variance includes only the variance components of the interactions of the target of measurement with the sources of error, i.e., those influencing one's relative standing with respect to others (Shavelson & Webb, 1991). In the current study, we included four variance components (i.e., $\sigma_o^2, \sigma_{r:s}^2, \sigma_{so}^2$, and $\sigma_{o(r:s),e}^2$) for absolute error variance and three variance components (i.e., $\sigma_{r:s}^2, \sigma_{so}^2$, and $\sigma_{o(r:s),e}^2$) for relative error variance. The observer variance was not included in the calculation of relative error variance because the main effect of observer was constant across sites and thus did not affect the rank ordering of sites.

Figure 1.
Reliability Coefficients across Observers and Rounds for the CIS and PPRS scales

Figure 2.
Reliability Coefficients for Different Combinations of Observers and Rounds for CIS-P and PPRS-LE

Table 1

Demographic Characteristics of the Programs (N=44)

Variable		% or Mean (SD)
Locale	Rural	11.4%
	Suburban	56.8%
	Urban	31.8%
Minority Enrollment	Low (< 25%)	36.4%
	Low-Medium (26–50%)	11.4%
	Medium-High (51–75%)	18.2%
	High (> 75%)	27.3%
Percent Free or reduced lunch Eligible Students	Low (< 25%)	25.0%
	Low-Medium (26–50%)	29.5%
	Medium-High (51–75%)	13.6%
	High (> 75%)	25.0%
Lowest Grade in Attendance	Kindergarten	64.3%
	1st Grade	21.4%
	2nd Grade	9.5%
Highest Grade in Attendance	4th Grade	2.4%
	5th Grade	76.2%
	6th Grade	11.9%
	7th Grade	4.8%
Grade of Enrolled Students		2.50 (1.70)
Total Number of Children Enrolled		24.24 (10.67)
Total Number of Staff		3.75 (1.75)

Table 2

Sample Descriptive Statistics for CIS (N=722)

Item	Mean	SD	CFA Result ¹	
			Factor Loading	Item R ²
Sensitivity (α =0.90)				
1 Speaks warmly to children	3.03	0.86	0.86	0.74
2 Listens attentively to children	3.02	0.85	0.84	0.71
3 Excited about teaching	2.76	0.90	0.85	0.72
4 Encourages children to try different experiences	1.86	0.91	0.55	0.30
5 Enthusiastic about children's activities/efforts	2.73	0.92	0.87	0.76
6 Encourages children to exhibit prosocial behavior	2.32	0.97	0.66	0.44
7 Positive physical contact	1.83	0.87	0.48	0.23
8 Sincere in tone of voice and manner	3.24	0.80	0.76	0.58
Harshness (α=0.79)				
9 Seems critical of children	1.16	0.46	0.79	0.62
10 Places high value on obedience-directive	1.68	0.87	0.51	0.26
11 Speaks with irritation or hostility to children	1.26	0.54	0.77	0.59
12 Negative physical contact	1.01	0.11	0.17	0.03
13 Punishes children w/o explanation or other redirection	1.10	0.36	0.58	0.34
14 Finds fault easily with children	1.12	0.42	0.86	0.74
15 Seems to prohibit many things children want to do	1.30	0.54	0.61	0.37
Detachment (α=0.79)				
16 Seems emotionally distant or detached from children	1.74	0.87	0.72	0.52
17 Spends considerable time not involving children	1.83	0.86	0.58	0.34
18 Seems interested in children's activities*	2.06	0.89	0.86	0.74
19 Supervises children closely*	2.09	0.90	0.62	0.38
Permissiveness (α=0.80)				
20 Exercises little control over children	1.55	0.72	0.29	0.08
21 reprimands children when they misbehave*	2.35	1.05	0.85	0.72

Item	Mean	SD	CFA Result ¹	
			Factor Loading	Item R ²
22	2.31	1.01	0.93	0.86
23	2.09	0.91	0.74	0.55

Note 1: The standardized factor loadings and item R² obtained from the Confirmatory Factor Analysis of the hypothesized 4-factor model.

Note 2: Items indicated by * are reverse-coded items.

Table 3

Sample Descriptive Statistics for PPRS (N=1092)

Item	Mean	SD	CFA Result ¹		
			Factor Loading	Item R ²	
Supportive Relations with Adults(SRA) ($\alpha=0.86$)					
1	Staff communicates clear expectations/guidelines	2.92	0.77	0.61	0.37
2	Positive behavior management techniques	2.94	0.82	0.70	0.49
3	Staff positive affect with students	3.02	0.77	0.82	0.68
4	Positive personal interactions with students	2.97	0.8	0.77	0.59
5	Positive engagement with students/activities	3.01	0.81	0.83	0.69
Supportive Relations with Peers(SRP) ($\alpha=0.89$)					
6	Positive peer interactions	3.18	0.69	0.86	0.73
7	Students work well together	3.21	0.70	0.90	0.82
8	Inclusive student atmosphere	3.18	0.73	0.81	0.65
Appropriate Structure (AS) ($\alpha=0.76$)					
9	Well managed activities	3.24	0.76	0.87	0.76
10	Clear and appropriate activity instructions	3.08	0.76	0.69	0.47
11	Staff cooperation and consistency	3.21	0.76	0.54	0.29
12	Safe student work environment	3.63	0.63	0.52	0.27
Level of Engagement (LE) ($\alpha=0.85$)					
13	Students positively engaged	3.24	0.71	0.87	0.76
14	Students on-task during activity	3.17	0.72	0.89	0.78
15	Students actively participating in on-task discussion	2.97	0.83	0.70	0.48
Chaos ($\alpha=0.83$)					
16	Inappropriate/disruptive student behavior	1.53	0.70	0.82	0.67
17	Ineffective student management skills/abilities by staff	1.48	0.71	0.87	0.75

Note 1: The standardized factor loadings and item R² obtained from the Confirmatory Factor Analysis of the hypothesized 5-factor model.

Table 4

Inter-rater reliability

Scale	Number of Items	ICC Reliability		
		Single Rater		Average of Two Raters
		(95% CI)		(95% CI)
<i>CIS</i>	23	0.63	(0.61 ~ 0.64)	0.77 (0.76 ~ 0.78)
Sensitivity	8	0.59	(0.57 ~ 0.62)	0.75 (0.72 ~ 0.76)
Harshness	7	0.42	(0.38 ~ 0.46)	0.59 (0.54 ~ 0.63)
Detachment	4	0.55	(0.50 ~ 0.59)	0.71 (0.67 ~ 0.74)
Permissiveness	4	0.45	(0.41 ~ 0.50)	0.63 (0.58 ~ 0.67)
<i>PPRS</i>	17	0.62	(0.61 ~ 0.64)	0.77 (0.76 ~ 0.78)
SRA	5	0.45	(0.42 ~ 0.49)	0.62 (0.59 ~ 0.65)
SRP	3	0.39	(0.34 ~ 0.44)	0.56 (0.51 ~ 0.61)
AS	4	0.52	(0.48 ~ 0.56)	0.69 (0.65 ~ 0.72)
LE	3	0.42	(0.38 ~ 0.47)	0.59 (0.55 ~ 0.64)
Chaos	2	0.51	(0.45 ~ 0.56)	0.67 (0.62 ~ 0.72)

Table 5

CIS Item Fit to the Revised 3-Factor Model (N=358)

Dimension/Item	Path Coeff.	Item R ²	Alpha ^a
<i>Sensitivity-Detachment</i>			0.92
Speaks warmly to children	0.88	0.78	
Listens attentively to children	0.86	0.74	
Excited about teaching	0.86	0.74	
Seems emotionally distant or detached from children	-0.69	0.47	
Encourages children to try different experiences	0.58	0.34	
Enthusiastic about children's activities/efforts	0.86	0.74	
Considerable time in activity not involving children	-0.52	0.28	
Positive physical contact	0.50	0.25	
Seems interested in children's activities*	-0.85	0.73	
Sincere in tone of voice and manner	0.78	0.60	
<i>Harshness</i>			0.83
Seems critical of children	0.75	0.56	
Places high value on obedience-directive	0.59	0.35	
Speaks with irritation or hostility to children	0.73	0.54	
Punishes children w/o explanation or other redirection	0.61	0.38	
Finds fault easily with children	0.88	0.78	
Seems to prohibit many things children want to do	0.70	0.49	
<i>Permissiveness</i>			0.87
Appropriately reprimands children when they misbehave*	0.86	0.74	
Exercises firmness when necessary*	0.94	0.89	
Expects children to exercise self-control*	0.70	0.49	

Note

^aReliability analysis was performed using the full sample (N=722)

Table 6

Variance Components Analysis Results (N=276)

Subscale	Mean	Var	Variance as a percent of total variance					Reliability coefficients	
			Site (σ_s^2)	Observer (σ_o^2)	Round:Site ($\sigma_{r:s}^2$)	Site×Observer (σ_{so}^2)	Observer×Round:Site + Residuals ($\sigma_{o(r:s),e}^2$)	D-index	G-coefficient
CIS-SD	2.75	0.34	35.8	9.6	22.9	0.0	28.7	0.63	0.68
CIS-H	1.30	0.14	42.8	6.0	20.4	11.9	22.8	0.67	0.69
CIS-P	2.16	0.51	13.1	17.6	13.6	17.7	35.8	0.33	0.39
PPRS-SRA	2.95	0.29	34.5	4.4	26.7	0.0	31.5	0.62	0.64
PPRS-SRP	3.17	0.27	17.0	11.7	31.5	3.7	30.0	0.38	0.42
PPRS-AS	3.27	0.23	23.9	1.1	21.1	14.6	37.8	0.50	0.51
PPRS-LE	3.11	0.28	10.6	4.5	39.0	13.1	27.8	0.25	0.26
PPRS-C	1.53	0.28	32.4	4.5	25.2	0.8	35.3	0.60	0.62