David P. Weikart Center for Youth Program Quality

Evaluation of Afterschool Improvement Process

Oklahoma 21st Century Community Learning Centers

March, 2017

the forum for youth investment

301 W. Michigan Ave. | Suite 200 | Ypsilanti, Michigan 48197 | Phone: 734.961.6900 | www.cypq.org

The David P. Weikart Center for Youth Program Quality is a division of The Forum for Youth Investment.

Evaluation of Afterschool Improvement Process

Oklahoma 21st Century Community Learning Centers

March 2017.

Charles Smith, Ph.D.
Leanne Roy
Steve Peck, Ph.D.
Gina McGovern
Katharine Helegda

Contents

Summary

Since 2007, the Oklahoma State Department of Education has operated a *quality improvement system* (QIS) for its approximately 100 federally-funded 21st Century Community Learning Centers (OK 21CCLC) afterschool programs with the explicit purpose of improving the performance of these service providers. This report draws upon data from 23 performance measures collected annually over multiple annual program cycles to present findings for reliability, validity, performance change, and effect of intervention fidelity on performance change. These analyses were conducted as part of an ongoing effort to: (a) evaluate over-time change in performance that is the central purpose of the QIS and (b) improve the accuracy and usefulness of performance data available to individual organizations that participate in the QIS.

In general, our findings indicate that the Oklahoma Afterschool Improvement Process is performing in accordance with its purposes: using accurate performance data to incentivize improvement in the quality of services.

Findings for the reliability and validity of the measures include:

- All of the 23 measures demonstrated acceptable levels of reliability.
- There is evidence for construct validity at each time point and factorial invariance across time points.

Findings for performance improvement include:

- Nearly all measures incrementally improved during a four year (2010-2013) period, while a subset demonstrated statistically significant growth.
- For nearly all measures, lower-performing sites at the baseline year (2010-2011) improved most. A subset of models demonstrated statistically significant effects.
- The indicator with the largest increase over four years was Targeting At-Risk Students, suggesting that even though the students served became more challenging, service quality was also generally improving.

Findings for intervention fidelity include:

- Higher fidelity of YPQI implementation is positively associated with growth on nearly all performance measures at over half of all year-to-year time increments, in line with the YPQI theory of change

This report is supplement to a series of annual reports submitted to the Oklahoma State Department of Education over eight years. These reports provide the unadjusted information that was used in the models described in this report. The supplement to the annual performance report for the 2013-14 program year (Sniegowski, Gersh, Smith, & Garner, 2015) provides the unadjusted means and descriptive statistics for all of the items and scales in the study.

Introduction

Since 2007, the Oklahoma State Department of Education has operated a *quality improvement system* (QIS) for its approximately 100 federally funded 21st Century Community Learning Centers (OK 21CCLC) afterschool programs with the explicit purpose of improving the performance of these service providers. Since 2009, the David P. Weikart Center has supported a QIS design with technical assistance and training, including supports for the *Youth Program Quality Intervention* (YPQI; (Smith, Akiva, Sugar, Lo, et al., 2012) [1] and the *Leading Indicators* suite of performance measures (Smith, Akiva, Sugar, & Hallman, 2012). The YPQI is an evidence-based intervention to improve the quality of instruction in afterschool programs, and the Leading Indicators are a suite of performance measures designed for afterschool systems that use an afterschool academic enrichment curriculum.

This report draws on data from these measures collected annually during multiple program cycles. This report advances the validity argument for these measures to a high standard of evidence by applying more sophisticated analytics (multi-level structural equation models) to questions about reliability, validity, and change. We believe that this standard for evidence should be required for use of performance measures in the public sector. These analyses were conducted as part of an effort to: (a) accurately describe aggregate changes in performance that the QIS was designed to produce and (b) maximize the accuracy of disaggregated data about individual organizations that are used in the QIS cycle.

Hypotheses

The QIS for the OK 21CCLC was designed to produce three types of change that should be visible on 23 of the performance measures that were tested using growth models. These types of change are:

- All features of 21st CCLC services should improve towards higher quality over time.
- Sites that start out with low quality should get turned around quickly to higher performance.
- Sites that implement YPQI at high fidelity will have larger performance gains compared to sites with lower fidelity.

Given these intended outcomes of the QIS, we fit data to growth models that reveal the extent to which (a) leading indicators tended to go up over time and (b) sites that started out low improved quality at a greater rate. In

[1] The YPQI is an evidence-based quality improvement intervention used in over four thousand agency, school, and community-based settings in 38 states. See also http://cypq.org/ypqi.

separate analyses, we tested relations between implementation fidelity and year-to-year change in performance measures.

Continuous Improvement Intervention

Quality Improvement Systems (QIS) in the afterschool field are designed to engage professional staff in the creation and implementation of high-quality services for both internal and external stakeholders using a lower-stakes accountability framework (Smith, 2013; Yohalem, Devaney, Smith, & Wilson-Ahlstrom, 2012). The Oklahoma QIS includes four core elements: (a) a set of standards for high-quality service and the aligned 21CCLC Leading Indicators suite of performance measures; (b) a continuous improvement cycle implemented annually at each program site; (c) performance reports for each program site and the overall system; and (d) training and technical assistance supports necessary to implement the continuous improvement cycle and performance measures.

The Oklahoma QIS is anchored by the *Youth Program Quality Intervention* (YPQI), an evidence-based continuous improvement cycle designed to embed a culture of continuous improvement at both the site and the network levels. This improvement culture is developed and maintained through a cycle of assessment, planning, and intentional improvement efforts that include the implementation of organizational practices that support the improvement process. Each site is expected to select a team to conduct several program self-assessments using the Youth Program Quality Assessment (Youth PQA; Smith & Hohmann, 2005). These assessments are conducted by individual observers rating a colleague's performance during a single *offering*[2]. After data are collected, site teams are expected to review the individual ratings and to submit an overall site assessment that represents the combined results of those ratings. In this way, site teams identify strengths and areas for improvement. A program improvement plan is then created based on these identified areas, and this plan includes detailed information about the timeline for the goals, parties responsible , resources and supports necessary, and a description of what success looks like. Throughout the program year, clients implement the steps necessary to achieve these goals.

These measures are supplemented with survey data collected from the major stakeholders associated with the QIS: Project directors, point-of-service[3] staff, participating youth, and parents are surveyed approximately midway through the spring of the programming year. Project directors and staff are surveyed about organizational and instructional practices as well as participation in and fidelity to the YPQI. Parents and youth are surveyed for their feelings about the effectiveness of the program. These surveys are known collectively as the Leading

[2] Offering is defined as a point-of-service setting where consistent groups of adults and youth meet over multiple sessions for the same learning purpose.
[3] Point-of-service is the setting level where adults deliver instruction to youth during program offerings.

Indicator Measures of Program Performance. The Leading Indicators[4] suite of performance measures was designed to align with performance standards for 21CCLC programs and are the source of performance information used in the YPQI. The Oklahoma design for QIS and YPQI is presented in Appendix A.

The Oklahoma State Department of Education (OSDE) also provides Technical Assistance (TA) coaches to select grantees. Grantees who receive coaching supports and services include all first-year grantees as well as those identified through the use of the Leading Indicator reports and recommended for services by OSDE. Two part-time coaches managed by the Weikart Center, and one coach managed by OSDE, provide comprehensive supports and services to up to fourteen 21CCLC grantees. Coaches work an average of 18 hours per week with the project director and staff of the assigned grantees. The coaching and TA services are designed to support grantees in their improvement process and include developing a TA Plan, providing support related to the program self-assessment process, facilitating regional training as requested, co-facilitating with the project director a data-planning session for program staff, modeling the observation-reflection method with a staff member, visiting program sites with the project director, and providing a year-end report that includes a summary of services, highlights, and recommendations for the future. Coaches also attend all state-wide or regional trainings.

Method

Participants

Performance data described in this report were collected each year from staff and students at OK 21CCLC program sites. Cohort sizes for the four years are reflected in Table 1. Appendix B provides additional descriptive information about the cohorts of staff and students in each year. Although parent measures are not examined in this report, we include parent-reported demographic information as an additional source of information about students.

Table 1. Cohort Sizes for Sites, Staff, and Students over Four Program Years

	Year 1	Year 2	Year 3	Year 4
Sites	107	107	86	99
Managers	98	114	121	109
Teachers	901	894	765	821
Students	3,485	2,990	2,464	2,781

[4] Several different Leading Indicators models are used in statewide 21st CCLC evaluations in Michigan, Washington, Arkansas, Oklahoma, Missouri, and New Jersey. The Michigan Leading Indicators used in this report were developed by evaluators at the Michigan State University Community Evaluation Research Collaborative (http://cerc.msu.edu/). Weikart Center's Leading Indicators measures are used in Arkansas and Oklahoma (Smith, Akiva, Sugar, & Hallman, 2012). Washington and New Jersey Leading Indicators systems are delivered by American Institutes for Research (http://www.k12.wa.us/21stCenturyLearning/Evaluations.aspx).

Performance Measures

The 23 performance measures that are the subject of this report are described in Table 2, which provides the measure name and a summary description of item content for each measure. Because the QIS/YPQI continuous improvement intervention is segmented by levels, the performance measures are designed to reflect the quality of services (internal or external) at multiple levels of organization. These "setting" measures are differentiated from individual student skills. There are: four measures at the system level; eight measures at the organization level; seven measures at the point-of-service level; and six measures of youth skill. Measures for each level reflect a mix of sources (e.g., manager, teacher, student), and these sources are reflected parenthetically in Table 2.

Table 2. Measure Names and Descriptions of Item Content for 23 Performance Measures

System Level of Setting	
Accountability	Program is held accountable for quality, is monitored routinely, and collaborates across sites; and staff, supervisors, and networks all use common standards of quality. (Manager)
Student Data	Program staff review test scores and grades from previous years and current school year, including diagnostic data. (Teacher)
Community Engagement	Students participate in civic engagement, and field trips/sessions are led or provided by local business groups or community groups. (Manager)
Organization Level of Setting	
YPQI Fidelity	Index of site manager responses regarding participation in planning, assessment, coaching, and training.
Staffing Model	Program staff arrive trained, receive program orientation, have adequate retention and staff/student ratios, are given time to plan, and have student goals in mind for program objectives. (Manager and Teacher)
Horizontal Communication	Staff co-plan program policies or activities with other staff, discuss problems, and observe or are observed by other staff. (Teacher)
Vertical Communication	Supervisor provides feedback, is visible during program, knows what is being accomplished, challenges staff, and makes sure program goals and priorities are clear. (Teacher)
Youth Program Governance	Youth are given opportunities to select content of activities, begin their own projects, and contribute to the aesthetics of the physical space. (Manager)
Youth Organization Governance	Youth are involved in hiring of new staff, allocation of budgeted funds, and help provide recognition of community volunteers and organizations that contribute to the afterschool program. (Manager)
Targeting Academic Risk	Students are targeted based on a below proficient score on local or state assessment, a failing grade during proceeding grading period, in need of additional assistance in math or reading, and English language Learner. (Manager and Teacher)
Point of Service Level of Setting	
Quality of Instruction	Staff employs practices supporting both exploratory learning and direct skill scaffolding. Practices further support generic skills necessary to manage emotions, motivation, and executive processes. (Observation)

Homework Completion	Youth feel like they are supported in understanding and completing their homework. (Students)
School Day Content	Program links parents with school day, including academic progress, and encourages participation in parent-teacher conferences. (Teacher)
Academic Planning	Program is targeted at specific goals and planned in advanced based on feedback from students to include content with the expressed interest of students. (Teacher)
Growth and Mastery Goals	Students are exposed to new experiences, have responsibilities and tasks that increase in complexity over time, work on long-term group projects, acknowledge achievements, and can identify personal strengths. (Teacher)
Youth Engagement	Youth feel challenged and interested at the program. (Student)
Youth Belonging	Youth feel like they belong and matter at the program. (Student)
Youth Skills and Beliefs	
Social Competencies	Youth works well with other kids, can make friends and stay friends, can talk to people they don't know, can tell other kids they are doing something they don't like, can tell a funny story, and can disagree. (Student)
Work Habits	Youth can follow rules, work well by themselves, are careful, make good use of their time, finish work on time, and can keep track of their things. (Student)
Reading/English Efficacy	Youth are interested in and good at reading/English, expect to do well, and would be good at learning something new in reading/English. (Student)
Math Efficacy	Youth are interested in and good at math, expect to do well, and would be good at learning something new in math. (Student)
Science Efficacy	Youth are interested in science and would be good at learning something new in science. (Student)
Technology Efficacy	Youth are interested in technology and would be good at learning something new in technology. (Student)

Staff and youth surveys were administered online via the online survey software Qualtrics unless a site specifically requested paper surveys. Each survey (online and paper) contained instructions for completing the survey as well as confidentiality assurances for youth. To ensure the protection of student confidentiality, all measures used have been reviewed and approved by an independent institutional review board (Chesapeake IRB) prior to the start of data collection in Oklahoma. All student-reported data provided in this report were aggregated to the site level for analysis and reporting.

Analytic Approach

In this report, we advance evidence for the reliability and validity of the Leading Indicator suite of performance measures to "industry standard" by applying more sophisticated analytics (e.g., multilevel structural equation models) to questions about reliability, validity, and change. We also test the multi-year performance data for each site using growth models to gauge the effectiveness of the Oklahoma QIS at raising the quality of services over four years. Our approach has been incremental, and Appendix C describes prior steps in this effort.

Two issues in the four years of performance data were of paramount concern: nested and missing data. First, we addressed nesting of the data; that is, students are nested within instructional settings that are nested within organizations that are nested within funding cohorts that are, finally, nested within the larger Oklahoma QIS. We estimated reliability and validity coefficients using statistical models that accounted for variance uniquely explained by the organization in which individuals were nested. We wanted to understand how reliable individual responses were and how reliable the site-level means were across individuals (i.e., how much agreement there was within each site). By using a random-effects model, we were able to fit the data to models that reflect the fact that instructional systems are embedded within organizational systems.

Second, our analytic approach to the growth trajectory models addressed systemically missing data, and our results should be understood in these terms. Specifically, OK 21CCLC has a rotating entry date for grants that last not more than five years. Our data included three cohorts, so a somewhat different group of sites was in the sample in each of the four years. We used models that accommodated this pattern of missing data to produce estimates of the quality of service available on each indicator, given the sites that were producing services in that year. This means that our growth results should be interpreted as representing the average amount of service quality available each year in a 21CCLC program site in the state of Oklahoma.

Finally, we wanted to test the YPQI theory of change by estimating the effect that high fidelity YPQI implementation has on performance over the full set of performance measures. In these analyses, we estimate general linear models (GLM), where the dependent variable (DV) was one of the 23 performance measures and covariates included the baseline for the DV, an indicator for site manager turnover, and the YPQI implementation index as a predictor. These analyses were conducted at a later date than those described above and so included six years of data between 2010-2011 and 2015-2016.

Results

Reliability

The scale reliabilities were assessed using Cronbach's alpha. Geldhof, Preacher, and Zyphur (2014) noted that common reliability statistics, including Cronbach's alpha, may become biased when applied to clustered data. The reason for this bias is that these reliability estimates conflate between-cluster variability and within-cluster

variability, causing the covariance matrix used to assess individual response consistency to also include between-group differences.

Geldhof et al. (2014) suggested addressing this problem by separately estimating a within-cluster (i.e., individual-level) and between-cluster (i.e., setting-level) reliability. The approach used for the staff and youth data was to remove between-cluster variability by group-mean centering all of the item data before calculating Alpha on the group-mean centered variables. The reliability results for the scales are displayed in Table 3.

Table 3. Cronbach's Alpha for 2012 and 2013 Adjusted for Between-Cluster Variability

	2012	2013
	Alpha	Alpha
System Level of Setting		
Accountability	.77	.66
School Day Content	.70	.76
Student Data	.73	.69
Community Engagement	.68	.84
Organization Level of Setting		
Staffing Model	.84	.81
Horizontal Communication	.85	.86
Vertical Communication	.88	.86
Job Satisfaction - Manager	.83	.86
Job Satisfaction - Teacher	.84	.84
Youth Program Governance	.65	.74
Youth Organizational Governance	.52	.68
Targeting Academic Risk	.83	.84
Point-of-Service Level of Setting		
School Day Content	.76	.79
Academic Planning	.82	.81
Growth and Mastery Goals	.86	.84
Homework Completion	.53	.51
Youth Engagement	.82	.80
Youth Belonging	.74	.70
Youth Skills and Beliefs		
Social Competencies	.72	.74
Work Habits	.78	.81
Reading/English Efficacy	.80	.83
Math Efficacy	.87	.88
Science Efficacy	.83	.86
Technology Efficacy	.81	.84

Construct Validity

There are two options for dealing with clustered data in traditional regression methods. The first option is to adjust the standard errors to account for differences in means and variances across clusters. This involves the application of robust standard errors, as is done for complex surveys that utilize clusters as the *primary sampling unit*. The second option is to explicitly model the multilevel structure using random effects. Both approaches were pursued but, due to challenges associated with the small sample of staff per afterschool site, the method using robust standard errors is reported in the Supplementary Technical Report (Albright, 2014).

We examined six confirmatory factor analysis (CFA) models that drew from multiple data sources (e.g., manager, teacher, student surveys; observations) with measures at a given level:

- Model 1 – System-level measures (13) from the manager data source
- Model 2 – Organization-level measures (21) from the manager data source
- Model 3 – Organization-level measures (14) from the teacher data source
- Model 4 – Point-of-service level measures (15) from the teacher data source
- Model 5 – Point-of-service level measures (11) from the student data source
- Model 6 – Individual-level student skill measures (23) from the student data source

The set of six models was estimated two times in the middle years of data in the series (i.e., 2012 and 2013), and the model fit statistics were used to indicate construct validity. Model fit statistics reflect the extent to which our theorized model (e.g., items mapped onto constructs according to theory) matches the observed data. Table 4 summarizes the results by reference to five common fit statistics associated with each of the six models:

- Chi-square – a chi-square statistic is computed to determine whether the covariance matrix produced by the expected model is significantly different from the covariance matrix produced by the observed data. In this case, a low and non-significant value is desired, suggesting that the two matrices are not different from each other.
- Root Mean Square Error of Approximation (RMSEA) – RMSEA is commonly reported as an alternative fit index to the Chi-Square. A lower value (usually less than .05) suggests a good fit, while a value of 0 indicates a perfect fit.
- Comparative Fit Index (CFI) – CFI ranges from 0 to 1, and larger value indicates a better fit. In general, a CFI greater than 0.90 implies adequate model fit, while a value of 1 indicates a perfect fit.

- Tucker-Lewis Index (TLI) – TLI is similar to CFI in that both indices are affected by the average correlations in the data. A value of 1 indicates a perfect fit, and a TLI greater than 0.90 implies adequate model fit.

- Standardized Root Mean Square Residual (SRMR) – SRMR is a standardized difference between the correlation matrix produced by the data and the correlation matrix produced by the model. A value of 0 indicates a perfect fit, although a value less than .08 is generally considered a good fit.

In Table 4, fit statistics falling in the "adequate" fit range are noted. Inferences vary from model to model, depending on which statistic is used but, for every model, there is at least one statistic that indicates an adequate fit and, in most cases, multiple statistics that indicate an adequate fit. We characterize this evidence of construct validity as moderate.

Table 4. Fit Statistics Summaries

Model	Year	Chi-Square	RMSEA	CFI	TLI	WRMR
Model 1	2012		Adequate	Adequate	Adequate	Adequate
	2013			Adequate		
Model 2	2012			Adequate	Adequate	
	2013			Adequate	Adequate	Adequate
Model 3	2012			Adequate	Adequate	
	2013			Adequate	Adequate	
Model 4	2012					
	2013				Adequate	
Model 5	2012		Adequate	Adequate	Adequate	
	2013		Adequate	Adequate	Adequate	
Model 6	2012		Adequate	Adequate	Adequate	
	2013		Adequate	Adequate	Adequate	

Invariance tests were used to examine the extent to which model fit was worsened where constraining the intercepts to be equal in two groups, with the groups here being the 2012 and 2013 samples. In most cases, the same model was supported for both years, meaning that the factor structure is the same at both points in time. This finding indicates that the quantitative growth estimates (e.g., intercepts and slopes) involving these points in time can be interpreted meaningfully (Widaman, Ferrer, & Conger, 2010).

Growth Trajectories

A growth curve analyses was performed over four years of data for each of the 23 performance measures. Full results are provided in the Supplemental Technical Report (Albright & Guyon-Harris, 2015). Table 5 shows that the fixed effect coefficient for year, representing the average slope across all sites, was positive for every Leading Indicator scale. Results for the following scales show that improvement occurred over time: targeting academic risk, accountability, job satisfaction, youth organization governance, and youth program governance. In each of

these cases, the confidence intervals around the mean in the first year of data do not overlap with the confidence intervals around the mean in the fourth year, indicating statistically significant changes. The results for two of the items (i.e., school day content and student data) revealed improvement over the first three points in time followed by lower scores in the final period. Very little change over time was observed for the staffing model variable.

The variance component for the random effect for year (i.e., the extent to which slopes varied by sites) was often near zero, meaning most schools showed similar trajectories. In some cases, particularly for the manager data, the negligible year variance component created problems during optimization; consequently, models were estimated using only a random intercept.

For each measure, a model was first attempted that included random effects for both the intercept (i.e., variation in school mean) and slope (i.e., variation in the rate of change from one school to the next). One benefit of including both random effects is to describe the extent to which starting points affect trajectories. A negative covariance between the intercept and slope means that schools starting off with low means tend to show the greatest improvement.

Table 5. Selected Coefficients for Growth Models

Measures	Fixed Effect Year	Random Effect Year	Random Intercept/ Slope Correlation
System Level of Setting			
Accountability	.13***	-	-
	(.023)		
School Day Content - Manager	.07	.05	-.05
	(.042)	(.047)	(.079)
Student Data	.07	.04	-.07
	(.042)	(.024)	(.044)
Organization Level of Setting			
Staffing Model	0.05	.03	-.07*
	(.031)	(.012)	(.027)
Horizontal Communication	.05	.02	-.03
	(.032)	(-.015)	(.027)
Vertical Communication	.05*	.01	-.04*
	(.022)	(.007)	(.017)
Job Satisfaction – Manager	.08*	.05	-.07*
	(.034)	(.016)	(.017)
Job Satisfaction - Teacher	.04	.01	-.01
	(.020)	(.006)	(.010)
Youth Program Governance	.130**	-	-
	(.043)		
Youth Organizational Governance	.21***	.01	.04*
	(.038)	(.009)	(.012)
Targeting Academic Risk	.26***	-	-

Measures	Fixed Effect Year	Random Effect Year	Random Intercept/ Slope Correlation
	(.038)		
Point-of-Service Level of Setting			
School Day Content - Teacher	.02	.02	-.04
	(.033)	(.017)	(.033)
Academic Planning	.04	.01	-.07
	(.028)	(.012)	(.023)
Growth and Mastery Goals	.05*	.00	-.02
	(.025)	(.011)	(.021)
Homework Completion	-.15	.01	-.00
	(.027)	(.009)	(.016)
Youth Engagement	-.16***	.00	.01*
	(.024)	(.003)	(.002)
Youth Belonging	-.12***	.00	-.01
	(.029)	(.010)	(.020)
Youth Skills and Beliefs			
Social Competencies	.03	.01	-.01
	(.021)	(.006)	(.009)
Work Habits	.04*	.01	-.01
	(.019)	(.005)	(.008)
Reading/ English Efficacy	.03	.01	-.01
	(.023)	(.007)	(.012)
Math Efficacy	.06*	.01	-.02
	(.024)	(.006)	(.014)
Science Efficacy	.05*	.00	-.01
	(.02)	(.001)	(.008)
Technology Efficacy	.07	.01	-.03
	(.024)	(.007)	(.015)

Note. Standard errors are in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$

Fidelity Effect

Because the YPQI is a complex intervention, fidelity includes numerous steps undertaken by multiple people at each afterschool site over the course of the school year. The intervention includes four primary parts:

- Team self-assessment of instructional quality
- Planning with Leading Indicators data
- Instructional coaching
- Training on specific instructional skills

We constructed an index of the four practices for each site, ranging between 0 and 4. This index was entered into a GLM after controlling for the baseline version of the DV and a staff stability indicator for each site. All models were run four times for a sequence of two-year models where the DV was the second year, the baseline was the

first year, and YPQI Fidelity index was the sum over both years. Effects sizes (standardized beta from GLM) for the YPQI Fidelity Index are presented for comparison in Table 6.

These models were estimated using a dataset that aligned all five-year 21st CCLC grants on a common start date. We then examined the relations between implementation fidelity and indicator change during all five-year grants for all years for which we have data. To summarize the results in Table 6, YPQI implementation was positively related to change on nearly all indicators in all years: 44 of 48 were positive, and 25 of 48 were statistically significant. No negative associations were statistically significance. As expected, the YPQI implementation was most strongly associated with quality improvement (positive change on an indicator) for indicators at the organization level, and implementation was strongly associated with growth in a collaborative culture.

Perhaps most importantly, YPQI implementation fidelity was positively associated with instructional quality at the point of service level. High YPQI fidelity is associated with more school day content alignment, more project-based learning experiences, and higher instructional quality (e.g., teacher supports for student management of emotions, motivation, and executive functions).

The YPQI theory of change suggests that YPQI implementation fidelity has its primary effect on organization and point-of-service level indicators – most importantly, improved quality of instruction. Given this theory, we expect the effect of YPQI fidelity on youth beliefs and behaviors to be weaker in linear models like those used here. Further, our lack of individually-identified data make these analyses less specific, and lacking statistical power. With these caveats in mind, we did conduct analyses using youth-level indicators (Table 2, bottom panel), Although most of the results at the youth level were non-significant, they were nearly all positive and indicated that sites with higher YPQI fidelity involved youth who were more engaged, got more homework done, and had more pro-social skills. More rigorous tests of fidelity effects on student-level outcomes will come in future analyses that more closely examine the relations between the point-of-service level indicators and change in student skills.

Table 6. Estimated YPQI Implementation Fidelity Effect on Indicator Change

	Year 1-2	Year 2-3	Year 3-4	Year 4-5
System Level				
Accountability	.43*	.23	.07	.11
Collaboration	.44*	.39*	.26	.20
Student Data	.09	.09	.37*	-.01
Organization Level				
Staff Model	.42*	.45*	.46*	.55*
Horizontal Communication	.79*	.33	.73*	.24
Vertical Communication	.86*	.45*	.74*	.39

Job Satisfaction	.46*	.36	.33*	.30
Youth Governance	.27	.67*	.28	-.05
Targeting	-.02	.20	-.12	.22
Point-of-Service Level				
Quality of Instruction	.08*	.07	.05	.05
School Day Content	.59*	.27	.52*	.27
Academic Planning	.90*	.22	.39*	.41*
Growth and Mastery	.89*	.56*	.54*	.58*

Note. Standardized betas are reported. * $p < .05$, ** $p < .01$, *** $p < .001$

Discussion

Since 2007, the Oklahoma State Department of Education has operated a *quality improvement system* (QIS) for its approximately 100 federally-funded 21st Century Community Learning Centers (OK 21CCLC) afterschool programs with the explicit purpose of improving the performance of these service providers. This report draws upon data from 23 performance measures collected annually over multiple annual program cycles to present findings for reliability, validity, performance change, and effect of intervention fidelity on performance change.

These analyses were conducted as part of an ongoing effort to: (a) evaluate over-time change in performance that is the central purpose of the QIS and (b) improve the accuracy and usefulness of performance data available to individual organizations that participate in the QIS. In general, our findings indicate that the Oklahoma Afterschool Improvement Process is performing in accordance with its purposes: using accurate performance data to incentivize improvement in the quality of services.

This report is part of a series of annual reports delivered over several years to the Oklahoma State Department of Education. The annual performance report for the 2013-14 program year (Sniegowski, Gersh, Smith, & Garner, 2015) provides the unadjusted means and descriptive statistics for all of the items and scales in the study. Analyses for this technical report were conducted by the Weikart Center and an analytics subcontractor.[5]

Because findings reported here indicate that most Oklahoma 21st CCLC programs are attaining a high level of service quality, including strong alignment to the school day and high levels of instructional quality, we suggest that it may be time to move the evaluation to a more direct focus on program effectiveness. In particular, we think that the question of how participation in very high-quality afterschool programs is related to school day achievement and behavior could be added as a focus of the evaluation.

[5] Supplementary technical discussions of methodology, analyses, and findings are provided in Albright (2014) and Albright and Guyon-Harris (2015)

Findings for Reliability and Validity

We consulted with a methodologist about the reliability analyses to assess the extent to which the scales included in the study were differentiating among both individuals and sites. Reliability of the site-level indicator suggests that the mean score over all of the individuals at a site could be accurately used to represent performance on that attribute for that site. All of the 23 measures demonstrated acceptable levels of reliability.

We conducted validity analyses using data from the 2011-2012 and 2012-2013 program years, replicating confirmatory factor analyses (CFA) and tests of factorial invariance to assess the extent to which the attributes that we were measuring (e.g., job satisfaction) were (a) reflected in the data and (b) interpreted similarly across time. Model fit statistics were in the desired range and suggested moderate support for our empirical specification of the 23 scales; that is, there is evidence of construct validity. The relations between the items and scales did not vary over two time points; that is, there was evidence of longitudinal factorial invariance of the measurement models. This indicates that the results of the four-year growth trajectory analyses can be meaningfully interpreted as reflecting, for example, change in the Leading Indicators rather than change in the measurement models.

Findings for Growth

The OK 21CCLC improved performance during the 2010-2013 period, across a wide range of indicators. While 21 of the 23 measures had positive coefficients, 10 of these were statistically significant indicating that positive growth occurred. 21st CCLC programs improved accountability systems, vertical communication, job satisfaction, youth program governance, youth organizational governance, targeting academic risk, growth and mastery goals, work habits, math efficacy, science efficacy. The indicator with the largest increase over four years was Targeting At-Risk Students, suggesting that even though the students served became more challenging, service quality was generally improving.

Two indicators, student engagement and student belonging were both negative and statistically significant. We believe that these downward trends may be associated with efforts to recruit more academically at risk students and increasing emphasis on alignment with school day content. By –year analyses indicate that the negative trend in these two indicator continued in the 2014-15 year and then rebounded slightly in 2015-16.

Almost all measures demonstrated the expected negative coefficients for the correlation between the intercept and slope. For five of the measures these coefficients were statistically significant indicating that lower scoring sites improved more over four years than sites with higher scores at baseline. This finding accords with the desire for fast turnaround for lower-performing sites.

Findings for Fidelity Effect

Higher fidelity of YPQI implementation is associated with significant positive change in a subsequent year on 50 percent of tests, in line with the YPQI theory of change. Ninety-two percent of these relations were positive. In general, higher YPQI fidelity was associated with a more collaborative organizational culture, stronger school day content alignment, more project-based learning experiences, and stronger instructional quality (e.g., teacher supports for student management of emotion, motivation, and executive skills).

Recommendations

In our efforts to evolve the evaluation of the Oklahoma Afterschool Improvement Process, we offer the following recommendations for consideration:

Consider redesign on Leading Indicators measures. Given the results in this report, we now have sufficient information to re-evaluate the Leading Indicators measures with a goal to reduce the overall number of items. We also suggest changes in the self-assessment protocol to improve precision of data collection and potentially add a feasibly brief measure of youth engagement.

Consider impact evaluation phase. There are several reasons to consider shifting the evaluation focus toward questions of impact of on student achievement. First, performance data from the Leading Indicators evaluation system suggests that the consistency and quality of OK 21CCLC services has improved in three areas: quality of instruction, academic press and connection to school day, and recruitment of academically at-risk students. These areas of instructional quality are likely related to school day success, and we know that many OK programs are delivering at high quality on these program features. Second, we now have access to individual student data as part of the routine evaluation activities, making use of that data for evaluative purposes lower cost.

Create a Total "Q" Rating for Sites. Finally, as a "reach" goal we believe that it is possible to use the OK Leading Indictors measures to construct an overall quality rating for each site, drawing upon all sources of data. This encompassing "Q" score would provide an easy metric to guide the afterschool improvement process. The proposed approach would employ a fully Bayesian model that (a) estimates latent variable scores on each scale within each data level and (b) uses the latent variable scores at each level to estimate an overall site score that accurately accounts for uncertainty in the different measurement models. The model would produce a single omnibus continuous quality score for each site along with a confidence interval around those estimates.

References

Albright, J. J. (2014). *Reliability and Construct Validity of Director, Staff, and Youth Data*. Ypsilanti, MI: Methods Consultants of Ann Arbor.

Albright, J. J., & Guyon-Harris, K. (2015). *Summary of Results from Growth Trajectory Analysis*. Ypsilanti, MI: Methods Consultants of Ann Arbor.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72.

Smith, C. (2013). *Moving the Needle on "Moving the Needle": Next Stage Technical Guidance for Performance Based Accountability Systems in the Expanded Learning Field with a Focus on Performance Levels for the Quality of Instructional Services*. Ypsilanti, MI: Weikart Center for Youth Program Quality.

Smith, C., Akiva, T., Sugar, S., & Hallman, S. (2012). *Leading indicators measurement system: Analysis of Oklahoma data - Technical appendix to the Oklahoma 21st Century Community Learning Centers statewide evaluation*. Ypsilanti, MI: Weikart Center for Youth Program Quality.

Smith, C., Akiva, T., Sugar, S., Lo, Y. J., Frank, K. A., Peck, S. C., & Cortina, K. S. (2012). *Continuous quality improvement in afterschool settings: Impact findings from the Youth Program Quality Intervention study*. Ypsilanti, MI: Weikart Center for Youth Program Quality.

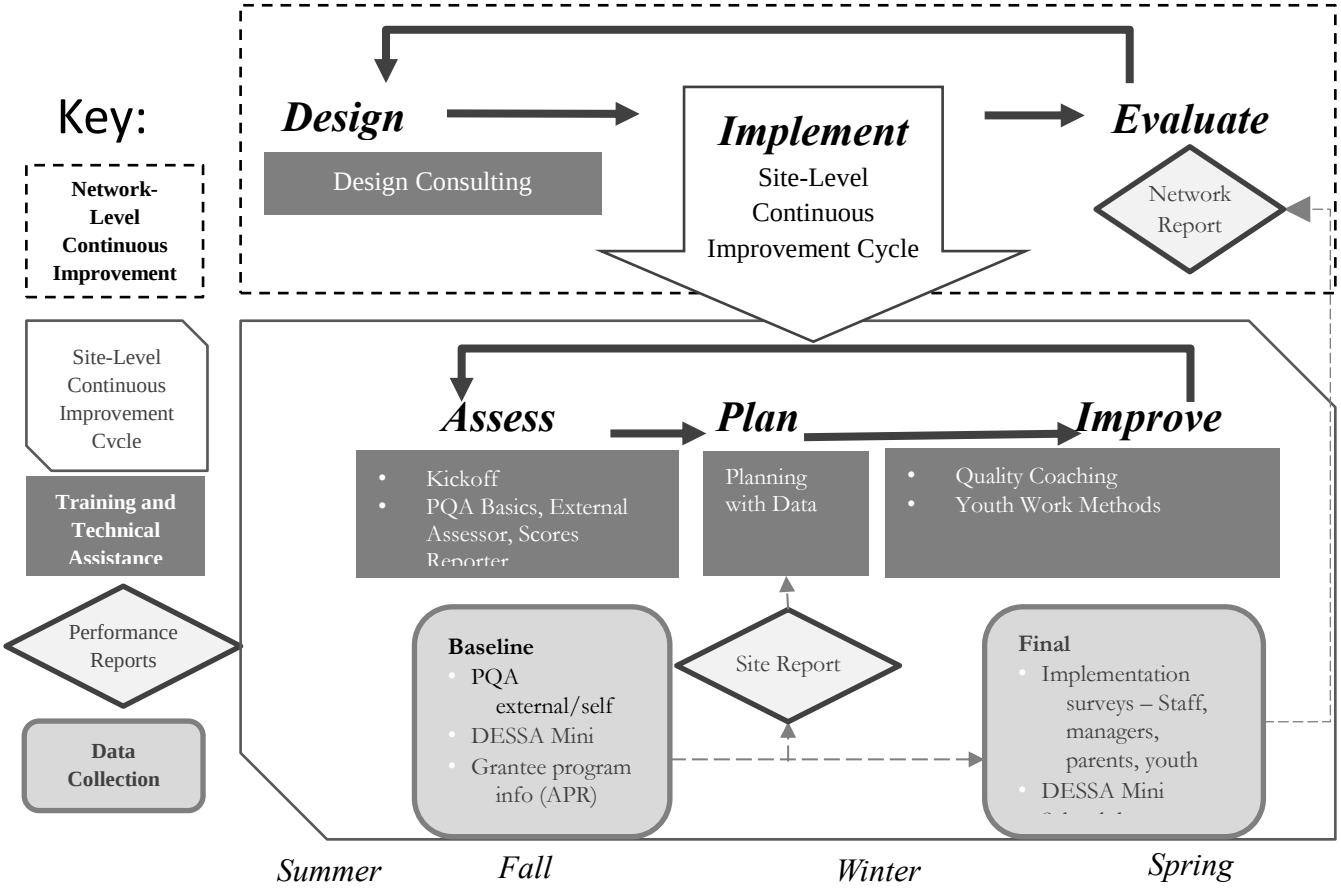
Smith, C., & Hohmann, C. (2005). *Full findings from the Youth PQA validation study*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Sniegowski, S., Gersh, A., Smith, C., & Garner, A. (2015). *Oklahoma 21st Century Community Learning Centers Statewide Evaluation Interim Report: 2013-2014 Annual Report*. Ypsilanti, MI: Weikart Center for Youth Program Quality.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10-18.

Yohalem, N., Devaney, E., Smith, C., & Wilson-Ahlstrom, A. (2012). *Building citywide systems for quality: A guide and case studies for afterschool leaders*. Washington, DC: The Forum for Youth Investment

Appendix A. Oklahoma QIS/YPQI Design



Design – Implement – Evaluate Cycle. The system-level continuous improvement cycle is the responsibility of system leaders and includes the design of the QIS, implementation of the assess-plan-improve sequence at program sites, and evaluation of aggregate performance information during each annual cycle and over multiple cycles.

Assess – Plan – Improve Cycle. This site-level continuous improvement cycle is typically the responsibility of the site manager and includes assessment of site performance, review of performance data, and implementation of an improvement plan for the site. This cycle is the Youth Program Quality Intervention (YPQI).

Training and Technical Assistance. These supports include a wide range of design consulting and evaluation services for the system-level cycle and the YPQI package of supports for program self-assessment of performance, planning with performance data, and implementation of the performance improvement plan.

Performance Data Collection. Performance data include the Leading Indicators. These data are the core of the planning element of the assess-plan-improve cycle.

Performance Reports. These data products are designed to support decision making at the site and system levels. Site-level performance reports are typically produced for each program site early in the annual cycle so that improvement planning can occur. The aggregate report, carrying information about all program sites, is typically produced after the end of an annual cycle so that design adjustments can occur during the system- and site-level cycles.

Appendix B. Characteristics of Staff and Students

Table B-1. Staff Characteristics

	(2011-2012) N=901	(2012-2013) N=894	(2013-2014) N=756	(2014-2015) N=821
Average years of experience at site	3	3.5	3.30	3.30
Education Level				
Less than high school diploma/GED	5%	4%	5%	7%
GED/High School diploma	9%	8%	9%	13%
Some college, no degree	11%	10%	11%	9%
Associate's Degree	3%	4%	4%	5%
Bachelor's Degree	40%	43%	44%	40%
Graduate program but no degree yet	8%	5%	6%	5%
Master's Degree	23%	24%	22%	19%
Doctorate	0.5%	1%	0%	1%
Other professional degree after BA	0.5%	0%	1%	1%
Teaching Certification	70%	69%	71%	63%
Average months worked per year	8.50	8.50	8.50	8.20
Average hours worked per week	7.80	8.50	8.34	8.35
Gender	10% male	12% male	14% male	12% male
Race				
White	85%	81%	80%	80%
African American	3%	3%	4%	4%
Native American	22%	20%	20%	21%
Hispanic	3%	3%	3%	4%
Arab American	0%	0%	0%	0%
Asian	0%	0%	1%	1%
Other Race	1%	39%	1%	1%

Table B-2. Youth Characteristics

Youth Survey	**(2011-2012) N=3,485**	**(2012-2013) N=2,990**	**(2013-2014) N=2,464**	**(2014-2015) N=2,781**
Average Age	11.5	11.6	11.70	11.67
Average Grade	5.70	5.80	5.78	5.77
Gender	49% male	48% male	50% male	50% male
Race (check all that apply)				
White	60%	58%	58%	60%
African American	10%	30%	35%	38%
Native American	35%	10%	9%	9%
Hispanic	11%	14%	14%	11%
Arab American	0%	4%	0%	1%
Asian	2%	2%	1%	1%
Other Race*	8%	7%	7%	7%

A total of 3,001 parents completed a survey, representing responses from 92% of Oklahoma 21st CCLC sites. Table B-3 displays information for the parent sample across four years of data collection.

Table B-3. Parent Characteristics

Characteristics	(2011-2012) N=2,679	(2012-2013) N=2,605	(2013-2014) N=2,752	(2014-2015) N=3,001
Average Age				
25 or less years old	4%	6%	6%	5%
26-30 years old	13%	18%	17%	19%
31-35 years old	23%	28%	26%	27%
36-40 years old	16%	20%	21%	19%
41-44years old	10%	12%	14%	14%
46-50 years old	7%	8%	6%	7%
51-55 years old	4%	4%	3%	4%
56-60 years old	2%	3%	3%	2%
61-65 years old	1%	2%	2%	2%
66 or more years old		1%	2%	1%
Education				
Less than high school diploma/GED	10%	12%	11%	11%
GED/High School diploma	24%	27%	31%	31%
Some college, no degree	20%	26%	26%	24%
Associate's Degree	9%	12%	11%	12%
Bachelor's Degree	13%	16%	14%	14%
Graduate program but no degree yet	1%	2%	1%	2%
Master's Degree	4%	4%	5%	5%
Doctorate	.3%	2%	1%	.5%
Other professional degree after BA	.2%	0%	0%	.5%
Race (check all that apply)				
White	62%	62%	60%	56%
African American	7%	6%	7%	8%
Native American	24%	24%	25%	28%
Hispanic	10%	10%	14%	8%
Arab American	0%	0%	0%	1%
Asian	2%	2%	1%	6%
Other Race	1%	1%	1%	1%
Gender	20% male	20% male	20% male	18.5% male
Income				
Less than $10,000	6%	8%	7%	8%
$10,000 to $19,999	11%	16%	17%	14%
$20,000 to $29,999	16%	18%	18%	21%
$30,000 to $39,999	13%	17%	17%	15%
$40,000 to $49,999	8.1%	11%	10%	10%
$50,000 to $59,999	8%	7%	7%	7%
$60,000 to $69,999	6%	6%	6%	5%
$70,000 to $79,999	4%	5%	5%	5%
$80,000 to $89,999	4%	4%	4%	4%
$90,000 to $100,000	3%	3%	4%	4%
More than $100,000	4%	5%	5%	5%

Appendix C. Prior Work on Measure Reliability and QIS Effectiveness

Each year, as part of the aggregate evaluation report, the Weikart Center conducts selected supplementary analyses designed to increase (a) the precision of performance information produced through the YPQI/Leading Indicators Evaluation measures and (b) what we know about the performance of the Oklahoma QIS. These aims correspond to two research questions:

- Are the performance data actually describing important attributes of the service? (e.g., Are the measures reliable and valid?)
- Do the data reflect expected patterns of performance improvement? (e.g., Is the QIS effective?)

These questions are critical to achieving the purposes of the YPQI/Leading Indicators intervention – to improve performance of Oklahoma 21st CCLC programs on the 23 performance indicators they have selected as representing effective services – thereby improving individual skill growth of student participants in the service. To this end, several supplementary analyses have been conducted since 2010:

- 2010-2011 – A separate technical appendix to the Statewide Evaluation report was focused on basic description of item performance, basic scale reliability, and convergent validity. See the report titled *Development and Early Validation Evidence for Leading Indicators Framework for Continuous Improvement in Afterschool Settings: Analysis of Oklahoma Data* (Smith et al., 2012).
    - Findings: Most scales had acceptable reliability; an expected pattern of positive correlations across the 23 measures was present (convergent validity); and higher-quality academically-related afterschool settings had higher reported participant satisfaction (staff, parents, and youth) and higher reported rates of homework completion.
- 2012-2013 – Two technical appendices were included in the *2012-13 Oklahoma 21st CCLC Statewide Evaluation* report (Sniegowski et al, 2014) that advanced the analyses of reliability a step further by examining the extent to which the average score for each performance measure was a reliable description of a site-level characteristic (Appendix A in the report). We also conducted basic analyses of change over multiple years using the *reliable change index* (Appendix B), a simple clinical method designed to assess how substantively important a change might be from one year to the next.

- Findings: In general, the mean score on each of the 23 performance measures was a precise description of site-level performance, with the exception of youth interest in academic subjects which we began to break out by different grade levels. Building on the reliability of the site mean-scores, we also learned that performance for a substantial number of sites was trending upward over the three years for which we had data: 2010-2011, 2011-2012, and 2012-2013. This was the first suggestion that substantively important improvement was occurring in specific sites on many of the 23 performance measures.

- 2013-2014 – We extended the prior analyses to include a *risk index* based on site performance across multiple performance measures (Appendix B).

    - Findings: Findings from the 2012-2013 report for reliability of site-level mean scores and growth were replicated. The risk index provided normative information; for example, 10 afterschool sites were low performing (i.e., in the lowest quartile of sample) on 10 or more performance indicators.