QTurn
Improvement science for the social and public sector

# Supplemental Impact Evaluation for Palm Beach County Quality Improvement System (QIS) using Fully Pattern-Centered Analytics

submitted to Lisa Lindeman
and Suzette Harvey

Jul 1, 2019

CONTENTS

## SUMMARY

Quality Improvement System (QIS) exposure moves afterschool programs to higher quality, increasing access to developmentally powerful settings and building children's social and emotional learning skills. Higher quality is defined in terms of the quality of instruction (i.e., individuation, basic/advanced SEL, enrichment content), the stability of staff tenure, and evidence of children's SEL skill growth.

In this study, we used performance data generated by Prime Time Inc. in Palm Beach County and fully pattern-centered methodology to describe the chain of causal effects as a cascade of sequential impacts. We sought to answer two specific questions about implementation and children's SEL skill growth: What is the impact of QIS exposure on program quality (i.e., best practices, low staff turnover, great content), particularly for programs that have lower program quality at baseline? What is the impact of exposure to high program quality on student SEL skills?

Findings demonstrate that (1) QIS exposure causes program quality improvement to occur and (2) exposure to high quality corresponds to SEL skill growth. Specifically, (1.a) quality increased dramatically over three years of exposure to the Palm Beach County QIS; (1.b) programs with Low Quality at QIS entry improved when exposed to even moderate QIS Fidelity; (2.a.) children exposed to higher-quality programs had greater SEL skill maintenance and gains compared to children exposed to lower-quality programs; and (2.b) children with Low SEL Skill at entry made greater gains at all levels of program quality.

This pattern of findings suggests that the Prime Time QIS design is successfully building the quality of services available in the county in substantively meaningful ways – by increasing the quality of instruction, increasing the tenure of staff, and growing SEL skills for students who need it most.

## CITATION

Smith, C. and Peck S. (2019). *Supplemental impact evaluation for Palm Beach County quality improvement system (QIS) using fully pattern-centered analytics*. QTurn LLC: Ypsilanti, MI.

## BACKGROUND

Quality Standard. This performance study defines the *quality of service* in afterschool programs in the terminology of child development. This means that program effectiveness is described in terms of the changes in children's skills and family/community situations. Children are at the center in this discussion of the quality of services for 190 afterschool providers in Palm Beach County, Fl. The County's long-standing commitment to developmentally-informed policies is evident in the definition of afterschool service quality that we at QTurn would characterize as having the right active ingredients: (1) using individuated instruction methods to build (2) basic and advanced social, emotional, and academic learning skills that (3) transfer to other settings, with a focus on (4) vulnerable children and youth. The ability to transfer skills to other settings is a powerful form of agency, and exposure to high-quality afterschool programs is an evidence-based approach to SEL-equity.

Evidence Base. The evidence base for the Palm Beach quality improvement system (QIS) is the deepest in the field, including studies by Chapin Hall, Weikart Center, and American Institutes for Research. This evidence based demonstrates staff support for the model, consistent QIS implementation, increasing quality of services, and positive impacts on children's social, emotional, and academic skills. Together, these studies provide a unique evidence base for the lower-stakes accountability approach to continuous quality improvement that Prime Time Inc. has pioneered. Importantly, the lower-stakes approach for adults has parallels to the definition of quality settings for children.

Questions and Design. In this study, we extend the evidence base addressing two broad research questions. How does exposure to the QIS impact the quality of services at each program, particularly programs that entered the QIS with very low service quality? How do social, emotional, and academic skills change when children are exposed to high-quality afterschool programs, particularly children who enter with lower skills?

The quality-outcomes (Q-O) design is a multi-level, pattern-centered, integrated evaluation design that yields rigorous estimates of program impacts using performance data that would be considered "underpowered" for more traditional research designs and methods. This study is a first application of a fully pattern-centered approach to impact analysis in the afterschool field. The Q-O design and pattern-centered methods are described in greater detail elsewhere (e.g., Smith et al., 2019; Smith et al., 2016).

The study sample includes 50 programs that entered the QIS in several cohorts since the 2013-2014 year. One of the requirements of the Q-O design is a *common baseline* for the QIS intervention – meaning that we moved the analytic frame of reference from calendar-year to

*QIS intervention-year* by treating all of the programs as if they started in the same year: QIS year one. From QIS year one, the data follow programs and students through QIS year three. This was necessary to model patterns of quality and impact that follow from exposure to the QIS. All findings for the study should be interpreted as effects produced by the *QIS three-year intervention design* rather than the effect of participation in the QIS for any of the three specific calendar years.

Inference and Generalizability. The Q-O study design provides a rigorous evaluation of the pattern of performance indicated in the data. There are several reasons why these data should be trusted for inferences about impact: First, the amount and quality of performance data generated by the Prime-Time systems is unique in the field. For example, few regional organizations have both complete information about staff participation in training and technical assistance (T&TA) and assurance that the T&TA is evidence-based and delivered at high fidelity. Further, the quality ratings produced through the QIS have sufficient precision (e.g., reliability) for use in quantitative models (Smith, 2013).

Another reason to trust the findings produced through the Q-O design is that the Prime Time theory of change allows for the application of "hammer-nail reasoning" – a type of research design that increases inferential power through the identification of a cascade of causal connections from (1) QIS entry and fidelity, through to (2) improvements in program quality and staff tenure, which in turn cause changes in (3) children's program-optimal skills and, finally, to change in (4) children's transfer-functional skills (Smith et al., 2019).

Pattern-centered methods are focused on identifying holistic types of performance that can have stability and/or can change shape over time. These findings are generalizable to other places and organizations that have similar types.

## QIS FIDELITY

Elements. The Palm Beach County QIS consists of four elements, all supported by Prime Time Palm Beach County, the County's nonprofit quality intermediary organization. This report uses indicator data for the first two elements: (1) receipt of continuous quality improvement (CQI) design supports and (2) participation in training, technical assistance, and networking events. Hereafter, the phrase *exposure to the QIS* refers to information about each of these two elements. Wage and scholarship incentives and expanded learning offerings will be considered in separate analyses.

1. **CQI design supports** – Includes supports for the evidence-based CQI design (i.e., the Youth Program Quality Intervention [YPQI]) and tools (e.g., PBC-PQA), coaching and project management supports, performance data collection and reporting, training calendars, etc.
2. **Training, technical assistance, and networking events** – Includes all of the training necessary to implement the CQI design (e.g., Self-assessment, Planning with Data), Youth Work Methods, Other trainings).
3. **Expanded Learning Offerings** – Includes programmed offerings from expert enrichment providers that Prime Time vets and selects based on quality.
4. **Wage and scholarship incentives** -

Indicators. Exposure to the QIS entails implementation of Prime Time's blended and sequenced adult learning design that integrates training, technical assistance, coaching, and networking events. The six indicators of QIS implementation fidelity are counts of manager and staff participation in three types of events: CQI training and technical assistance events (CQI events), Youth Work Methods training events (YWM event), Other professional development training or networking events (OTN event). Each event is approximately 0.5 days of staff time.

In order to better understand the meaning of attendance at these events, Prime Time expert practitioners helped set *quasi-absolute scales* (Bergman et al., 2003) for each of the six performance indicators where: 0 = none (i.e., no QIS fidelity), 1 = minimal (i.e., minimal QIS fidelity), 2 = moderate (i.e., moderate QIS fidelity), and 3 = high (i.e., high QIS fidelity).

Table 1 presents the minimum number of events per year for each type of event. The first number in each cell describes the minimum number of events necessary to be rated High QIS Fidelity using the quasi-absolute scale. The second number in each cell describes the minimum number of events necessary to produce a tangible effect on the program – which is the lowest performance possible in Moderate QIS fidelity. The following key performance indicators (KPIs) were developed:

- To achieve high QIS fidelity, managers should attend a total of 18 QIS events over three years, or an average of 6 events per year, and staff should attend 9 QIS events over three years, or 3 events per year.
- Minimal implementation – the point at or below which little effect is expected – is defined as 2 QIS events per year for managers and 1 event per year for staff.

Table 1. QIS Fidelity KPIs: Manager and Staff Minimum Number of Events Over Three Years for High and Moderate QIS Fidelity

| High/Low | CQI Event | YWM Event | OTN Event (other PD) | OTN Event (network) | Total | Annual |
|---|---|---|---|---|---|---|
| **Manager** | 2.5/1.0 | 3.0/1.0 | 2.5/1.0 | 6.0/1.0 | 14.0/4.0 | 4.6/1.7 |
| **Staff** | 1.3/0.5 | 3.0/1.0 | 1.3/0.3 | 1.0/0.25 | 6.6/2.1 | 2.2/0.7 |

Profiles. The six indicators of QIS fidelity are counts of manager and staff participation in three types of event - CQI, YWM, and OTN – composed as an average of averages. First, counts of events for managers and for staff are averaged within programs to create program-level event participation scores for mangers and for staff. Then, these two scores are averaged within each program (equally weighted) to produce a single event participation score for each program.

Figure 1 presents a summary of QIS fidelity in the sample by collapsing across 10 profiles. The full 10 profile solution – describing the actual typology of exposure to the QIS that exists in the County – is presented as Figure 2. From the information summarized in Figure 1, twenty two percent of QIS programs achieved High QIS Fidelity over the first three years of participation, whereas fourteen percent fell below the minimal-implementation KPI. A large majority, sixty four percent, achieved Moderate QIS Fidelity. From the information summarized in Figure 2, the ten types of QIS responses indicate that Prime Time helps afterschool program leaders customize their suite of QIS supports and achieve responsiveness to differing needs.

Figure 1. Summary of QIS Fidelity Profiles

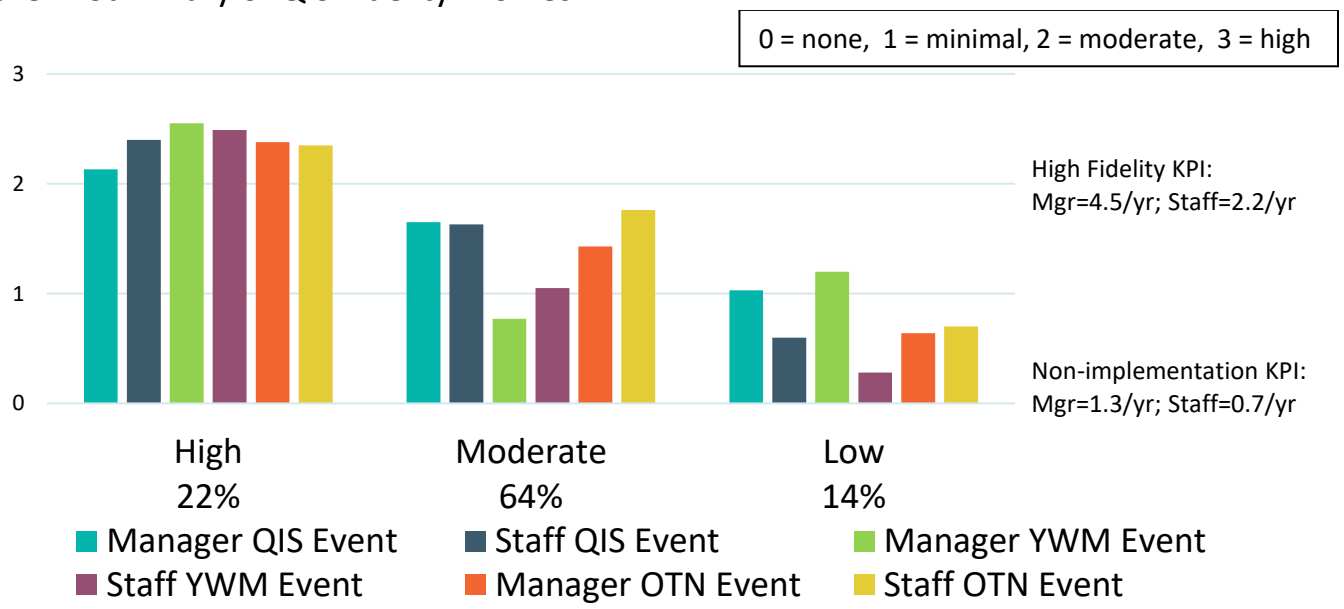
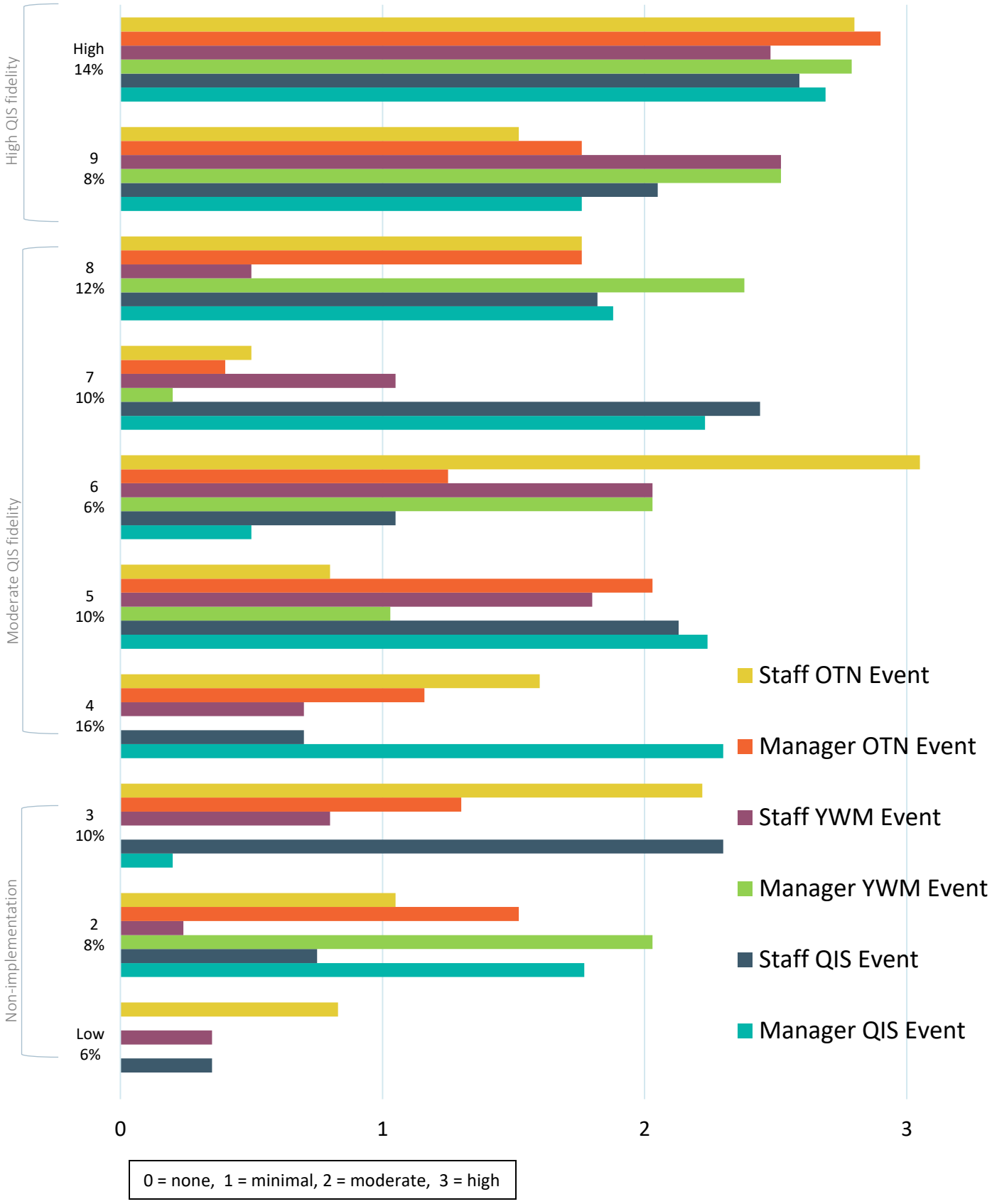


Figure 2. Complete set of QIS Fidelity Profiles



## QIS IMPACT

According to the Prime Time theory of change, exposure to the QIS elements will produce a cascade of positive effects on program quality, staff turnover, and student skills.

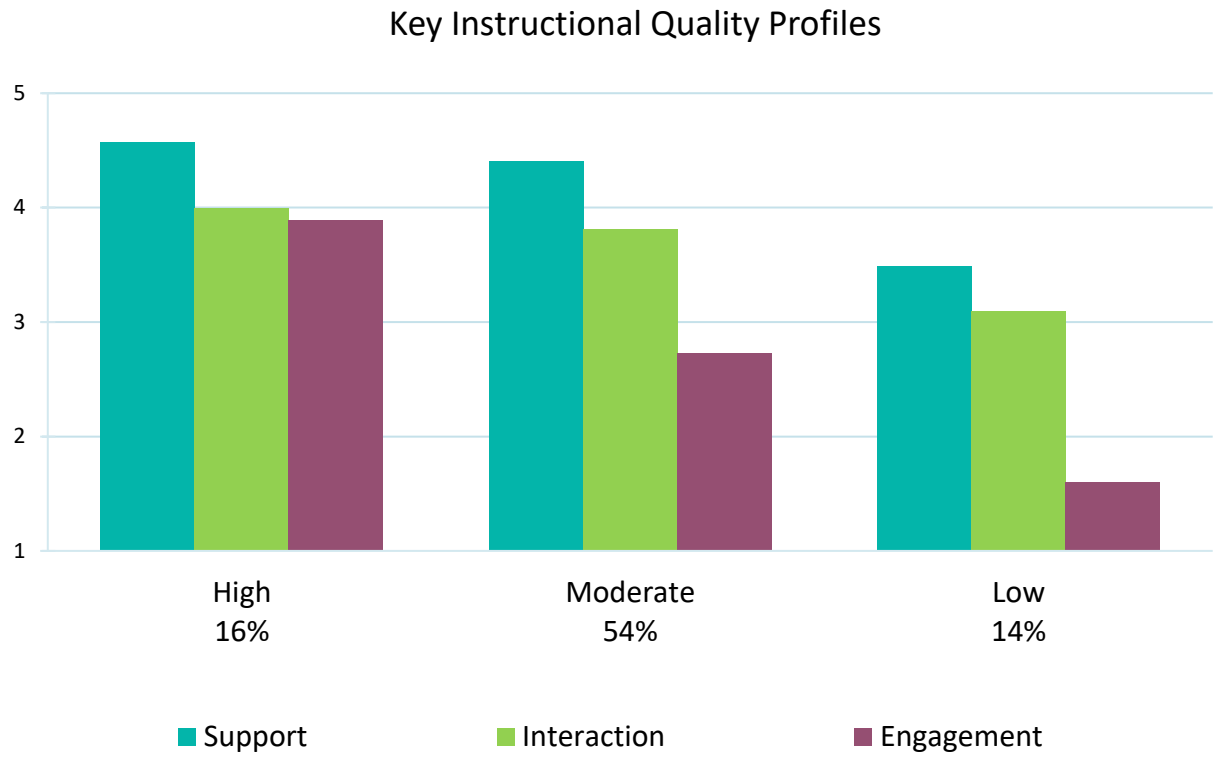
### Program Quality.

The results indicate substantively important increases in the quality of instructional practices by the second year of exposure to the QIS, and these increases were largely maintained or increased further by QIS year three. Over the first three years of exposure, seventy-two percent of programs improved program quality, and 51% of programs received a quality rating of 4.1 or more – the KPI for high-quality programs in the QIS. This was compared to 20% at or above the KPI at QIS entry. A detailed description of change in program quality for the first three years of exposure to the QIS using fully patter-centered analytics is provided in Appendix A.

Figure 3 describes four types of program quality available in the county during the period 2013-2018:

- *Very High Quality* was a new exemplary program type that emerged only after exposure to the QIS. Increasing availability of his highly responsive program design is likely to produce *SEL equity* effects which are discussed in the Student Skills section below.
- *High Quality* was a type of quality available in the county at baseline but the proportion of sites at or above the High Quality level grew from 11% to 50%.
- *Moderate Quality* (as shown in Figure 3) is not an empirically defined profile but the combination of ("collapsing over") multiple middling profiles for purposes of summarization.
- *Low Quality* was eliminated from the programs in this sample following exposure to the QIS; that is, low quality occurs only outside of the QIS in Palm Beach County after at least one year of QIS exposure. Low quality programs – those that fail to implement basic safety and youth engagement – can put vulnerable students at greater risk, creating SEL inequity.

Figure 3. Key Program Quality Types across 3 QIS Years




*Were these changes caused by exposure to the QIS?*

Table 2 compares patterns of change in Program Quality by levels of QIS Fidelity. For programs exposed to High QIS Fidelity, ninety-one percent of programs improved Program Quality whereas fifty-four percent advanced two quality levels or more. In contrast, for programs that were exposed to minimal-implementation QIS, twenty-nine percent improved whereas fourteen percent advanced two quality levels or more. For programs that had Low Quality at baseline: Fifty percent improved when exposed to at least Moderate QIS Fidelity, whereas zero percent improved when exposed to minimal-implementation QIS.

Table 2. QIS Fidelity and Quality Improvement over QIS Years 1-3

| | **Moved to any higher Q profile from baseline (N=50)** | **Moved to 2-higher Q profile from baseline (N=50)** | **Changed to any higher Q cluster – Only low Q from baseline (N=20)** |
|---|---|---|---|
| High QIS Fidelity | 91%* | 63% | 33% |
| Moderate QIS Fidelity | 75% | 59% | 69%** |
| No or Minimal-Implementation | 29%** | 14%** | 0%** |

Turnover. Turnover was defined at the program level in terms of the percent of program staff who started their job in the program during the current QIS cycle. The percent of turnover during each of the first three years of exposure to the QIS was treated separately in some analyses and averaged across years (mean = 24%) for other analyses. In addition, for some analyses, the average turnover across years was dichotomized such that programs with less than 30% turnover were categorized as having had no turnover, and programs with 30% or more turnover were classified as having had turnover.

*Were these changes caused by exposure to the QIS?*

Exposure to the QIS corresponded to reduced within-year turnover and buffered against negative effects of turnover on quality. However, this effect was present only for programs exposed to Moderate or High QIS Fidelity. All programs exposed to no or minimal QIS implementation and high staff turnover failed to improve service quality over three years.

As described in Table 3, exposure produced within-year staff turnover reductions from thirty-six percent of staff on average across sites to twenty percent. Turnover was reduced from forty percent of staff on average, across low baseline sites, to twenty-two percent.

Table 3. Average Staff Turnover in QIS Year 1 and 3 by QIS Fidelity and Program Risk

| | **All Programs N=26** | **High QIS Fidelity N=9** | **Mod QIS Fidelity N=15** | **Minimal implement N=2** | | **Low Quality N=11** |
|---|---|---|---|---|---|---|
| **QIS Year 1** | 36% | 43% | 33% | 25% | | 41% |
| **QIS Year 3** | 20% ($p < .05$) | 17% ($p < .2$) | 22% ($p < .1$) | 17% | | 22% ($p < .2$) |

Child Skills. According to Prime Time's QIS theory of change, exposure to high-quality afterschool programs – programs meeting the County's standard for quality – causes gains in child SEL skills, particularly for the vulnerable children who enter afterschool programs with lower SEL skills. Because low quality was eliminated in the QIS, we use the lowest moderate range quality score, Mod-Low, to provide the greatest contrast. However, the absence of a low-quality profile (i.e., low on the PQA scale is Safety<4, Supportive Environment and Interaction<3, Engagement<2) in the QIS, means that the contrast is likely to be weaker than in systems where low quality continues to be an issue.

Table 4 presents details on the number of children in the sample and the distribution of higher-risk children (i.e., low SEL skills at pre-test) across levels of Program Quality available in the QIS. Over three years, the total number of children in the study with complete data (i.e.,

program quality data and both pre and post SEL measures) was 1,758. The number of afterschool programs contributing these student ratings ranged from 21 to 26 in each of the three years.

The percent of youth with low SEL skills at entry was consistent over three years, although lower-quality programs had substantially more children entering with low SEL skills – a common pattern in the afterschool field. Across programs at the three levels of quality, the percent of children with low SEL skill at entry ranged between 20% and 48%.

Table 4. SEL Skill Calendar-Year Change Sample and SEL Risk Distribution

| | 2015-2016 | 2016-2017 | 2017-2018 |
|---|---|---|---|
| # pre-post students | 700 | 650 | 408 |
| # sites | 26 | 26 | 21 |
| % of all youth with low SEL skills at entry | 34% | 32% | 32% |
| % low SEL skills in High Quality Programs | 20% | 31% | 31% |
| % low SEL skills in Moderate Quality Programs | 34% | 33% | 35% |
| % low SEL skills in Mod-Low Quality Programs | 48% | 31% | 40% |
| % of sites with more than 60% low SEL skills | 12% | 8% | 10% |

*Did children's SEL skills increase following exposure to QIS programs?*

Tables 5, 6, and 7 describe correspondence between exposure to either the QIS's highest quality programs (i.e., High Quality) or the QIS's lowest quality programs (i.e., Mod-Low quality) in three successive years. We examined the Quality-Outcomes relationship – between exposure to program quality and direction of skill change – for children who entered programs at higher, more moderate, and lower SEL skill levels.

Of the three pairs of estimates that allow for comparison, two indicate the expected pattern, with higher-quality settings producing better outcomes. In general, high-quality settings retain more children at high levels of SEL skill. For one of the years, children with lower SEL skills had more skill gain where exposed to high compared to lower quality (2016-2017), whereas in the other two years children with lower SEL skills had more skill gain where exposed to lower compared to high quality (2015-2016 & 2017-2018).

Perhaps the most important result of these analyses are the results for lower SEL skill children. Regardless of exposure to High or Mod-Low quality, students who entered a program year with lower SEL skills made gains at twice the rate of children who started at the moderate level – on average 59% of lower SEL children made positive gains over the afterschool program year compared to 29% of children with SEL skills in the moderate range. Children with lower SEL skills gain more in Palm Beach QIS programs compared to their moderately skilled peers.

Further, high-quality programs appear to create conditions that help students maintain high levels of skill during the program.

Table 5. 2015-2016 DESSA – SEL Skill Change by Quality Level

| | High SEL Skill N=173 | Moderate SEL Skill N= 293 | Low SEL Skill (At-risk) N=235 |
|---|---|---|---|
| High Quality | 71% stable high*** | 14% gain | 38% gain*** |
| Mod-Low Quality | 55% stable high*** | 26% gain | 46% gain*** |

Table 6. 2016-2017 DESSA – SEL Skill Change by Quality Level

| | High SEL Skill N=161 | Moderate SEL Skill N= 419 | Low SEL Skill (At-risk) N=70 |
|---|---|---|---|
| High Quality | 62% stable high*** | 25% gain | 72%gain*** |
| Mod-Low Quality | 56% stable high*** | 37% gain | 62% gain*** |

Table 7. 2017-2018 SRYB – SEL Skill Change by Quality Level

| | High SEL Skill N=38 | Moderate SEL Skill N=201 | Low SEL Skill (At-risk) N=169 |
|---|---|---|---|
| High Quality | 72% stable high*** | 44% gain* | 59% gain*** |
| Mod-Low Quality | 40% stable high*** | 26% gain | 77% gain*** |

## CONCLUSIONS

Summary of findings. Quality Improvement System (QIS) exposure moves afterschool programs to higher quality, increasing access to developmentally powerful settings, and building children's social and emotional learning skills. Higher quality is defined in terms of the quality of instruction (i.e., individuation, basic/advanced SEL, enrichment content), the stability of staff tenure, and evidence of children's SEL skill growth.

In this study, we used performance data generated by Prime Time Inc. in Palm Beach County and fully pattern-centered methodology to describe the chain of causal effects as a cascade of sequential impacts. We sought to answer two specific questions about implementation and children's SEL skill growth: What is the impact of QIS exposure on program quality (i.e., best practices, low staff turnover, great content), particularly programs that have lower program quality at baseline? What is the impact of exposure to high program quality on student SEL skills?

Findings demonstrate that (1) QIS exposure causes program quality improvement to occur and (2) exposure to high quality corresponds to SEL skill growth. Specifically:

- *Quality Increased dramatically over three years of Exposure to the Palm Beach County QIS.* For all programs exposed to the QIS over three years:
    - 355% increase in High Quality, including emergence of a new exemplary program type (Very High Quality) - and elimination of the lowest quality type.
    - 44% reduction to within-year staff turnover.
    - 100% of programs exposed to Minimal QIS Fidelity and high staff turnover failed to improve over three years.
- *Programs with Low Quality at QIS entry improved when exposed to even moderate QIS Fidelity:*
    - 50% improved when exposed to at least Moderate QIS Fidelity while 0% improved when exposed to Minimal QIS Fidelity.
    - 43% reduction to within-year staff turnover.
- *Children exposed to higher-quality programs had greater SEL skill maintenance and gains* compared to children exposed to lower-quality programs.
- *Children with Low SEL Skill at entry made greater gains at all levels of program quality*. 103% more children with Low SEL Skill at program entry made gains compared to children at Moderate levels of SEL skill.

Recommended Next Steps. Consideration of next steps includes two components: Recommendations for extension of the present performance evaluation and recommendations for improving Prime Time information systems and practices.

### *What can we still learn from the Palm Beach data?*

The Palm Beach afterschool data is produced through some of the most sophisticated and complete data collection infrastructure in the afterschool field; that is, exemplary data systems. For this reason, Prime Time has the most complete information about professional development and job turnover and succession, and program quality, that we have ever encountered. This means that questions of basic scientific and policy relevance can be addressed using these data – we are just scratching the surface in this report. We recommend going to an external funder (e.g., Robert Wood Johnson) to secure more funding for questions of relevance to the broader field.

However, several smaller research questions may be of more immediate value to Prime Time:

- *What is the distribution of low SEL skill at entry (i.e., risk) across all programs, and do programs with very different levels of risk need different supports from the QIS?* For example, in Table 4, for the year 2016-17, 32% of all children were in the low SEL skill

group at entry. However, it is much more important to know the proportion of low SEL skill children for each individual program. For example, 12%, 23%, and 14% of the programs (for 2015-2016, 2016-2017, and 2017-2018, respectively) had less than 15% of children with lower SEL skills, whereas 12%, 8%, and 10% of the programs had more than 60% of children with lower SEL skills at entry. These are programs with very different types of need from the QIS.

- *How do the changes in program quality and child skill compare to other exemplary programs?* Determine Cohen's-d type effect sizes for program quality change and SEL skill change and compare to effect sizes in other studies using the same measures.
- *Does lower staff turnover correspond to greater SEL skill change?* For example in 2016-2017, of the children who improved SEL skills, 24% were in programs with High Turnover and 76% were in programs with Low Turnover.
- *Do academically at-risk students who are exposed to high quality and have large SEL skill gains transfer those skills to school as improved outcomes?* Data for school outcomes in the current district-provided data files is insufficiently matched to the existing Palm Beach data to support analyses. Specifically, a student sample matched to the programs for which Prime Time has both program quality and pre-post SEL skill measures would be required to substantially advance our understanding of afterschool impacts on school-day performance.

We also offer recommendations for Prime Time's next steps in building 2.0 knowledge management systems. This supplementary report and the broader performance report produced by Lindeman (2019) describe performance for the Palm Beach QIS at a level of detail and precision which is unique in the field. These reports describe operation of afterschool organizational processes at a sufficiently granular level – QIS fidelity, program quality, SEL skill change - to support reasoning about how to improve and know if it is happening or not.

One of the most important parts about the study is the creation of valid norms for performance. We now actually can empirically define what levels of QIS fidelity should be achieved to move program quality (See Table 1). We have also empirically defined what "high quality" looks like (See Figure 3) in Palm Beach County afterschool programs. However, the sample of children for SEL assessment was drawn from only one band of quality – high quality – and had almost no overlap with the school district data that was available to us. This prevented us from developing a full set of norms for SEL skill change, and ultimately, afterschool effects on school outcomes. A future round of SEL ratings could be conducted to produce more complete set of SEL skill change norms – and evaluating the association between quality, turnover, and child SEL skill change that are hinted at in the current study. \

While norms for performance and outcomes are a critical part of the QIS technology – the yardstick for telling us what's possible and what can be improved – the hardware and software

are also important.  Prime Time's sophisticated information management architecture – Sales Force, Scores Reporter, Registry, On-line surveys, School Data, Coaching records, etc. – has just been forced to into a state of integration by the Prime Time research questions and Prime Time staff's sophistication with R-code.  These staff's recommendations about how to improve the Prime Time information architecture are both worth tens of thousands of dollars in consulting time and leading edge, i.e., it would be hard to find a consultant that could even do the work.

## References


Bergman, L. R., Magnusson, D., & El-Khouri, B. M. (2003). *Studying individual development in an interindividual context: A person-oriented approach*. Mahwah, NJ: Erlbaum.

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., McGovern, G., Peck, S. C., Larson, R. W., Hillaker, B., & Roy, L. (2016). *Preparing youth to thrive: Methodology and findings from the social and emotional learning challenge*. Washington, DC: Forum for Youth Investment.

Smith, C., Peck, S. C., Roy, L., & Smith, L. (2019, April). Measure once, cut twice: Using data for continuous improvement and impact evaluation in summer programs. In D. Darling (Chair), *Digging deeper: Using mixed-methods research to assess promising practices in out-of-school time*. Symposium conducted at the annual meeting of the American Educational Research Association, Toronto, ON.

# APPENDIX A

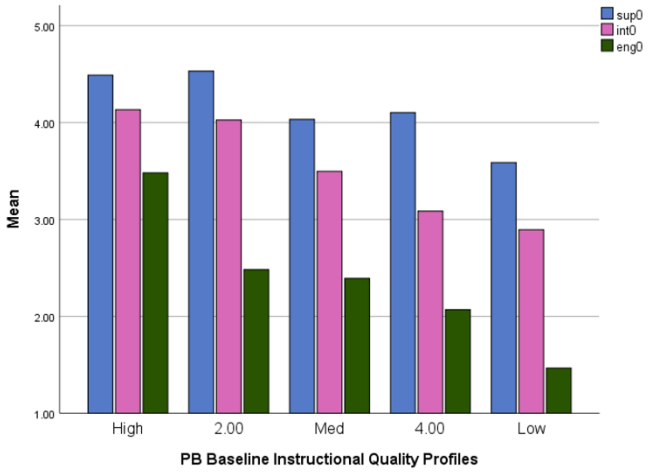
## Structural (aka, Sample-level) Stability and Change in PQA Profiles

According to Bergman, Magnusson, and El-Khouri (2003), a central task of the pattern-centered analysis of change in multivariate profiles over time is the examination of "the evolvement of the classification structure (structural stability and change)" (p. 47); that is, "in studying structural stability, we compare two sets of centroids [i.e., the set of means on the cluster variables within each cluster group] taken from different measurement occasions" (p. 152). The purpose of examining structural stability and change is to determine the extent to which "clusters in one classification at one time are reproduced in the other classification at the next time" (p. 123). In other words, we need to determine the extent to which the profile shapes found at one point in time match the profile shapes found at another point in time, which includes identifying profiles at the first time point that do not appear at the second time point and profiles at the second time point that did not appear at the first time point.

We sometimes refer to structural stability and change as *sample-level* stability and change because we are trying to identify the number and form of the most common profile shapes characterizing a given sample, both at each point in time and, collectively, across all points in time. For example, we might find five profiles at Time 1 and seven profiles at Time 2, and this would indicate structural change. We might also find five profiles at both Time 1 and Time 2, but the shapes of some or all of the profiles might be different at the two points in time, and this would also indicate structural change across time. If we found five profiles at Time 1 and seven profiles at Time 2, and four of the Time 1 profiles were identical to four of the Time 2 profiles, we could describe this pattern as "*partial structural stability*" (Bergman et al., 2003, p. 136) or substantial structural change.

Further, we distinguish sample-level stability and change from individual-level stability and change because they reflect two completely different phenomena. For example, even if we found perfect structural stability across two points in time (i.e., the number and shapes of the profiles were identical across time), the specific individuals (or programs) characterized by a given profile shape at Time 1 (e.g., Profile X) can be completely different than the specific individuals (or programs) characterized by the identical profile shape at Time 2. This would occur if all of the individuals in Profile X at Time 1 moved to Profile Y at Time 2, and all of the individuals who moved to Profile X at Time 2 moved from profiles other than Profile X at Time 1. Although this kind of extreme combination of structural stability and individual-level change over time is not common, it is quite common to find large percentages of individuals moving into and out of similar profiles across time. In order to understand the meaning of these kinds of complex combinations of stability and of change, we first need to understand the full range and forms of sample-level profile shapes across the entire period of time being studied.

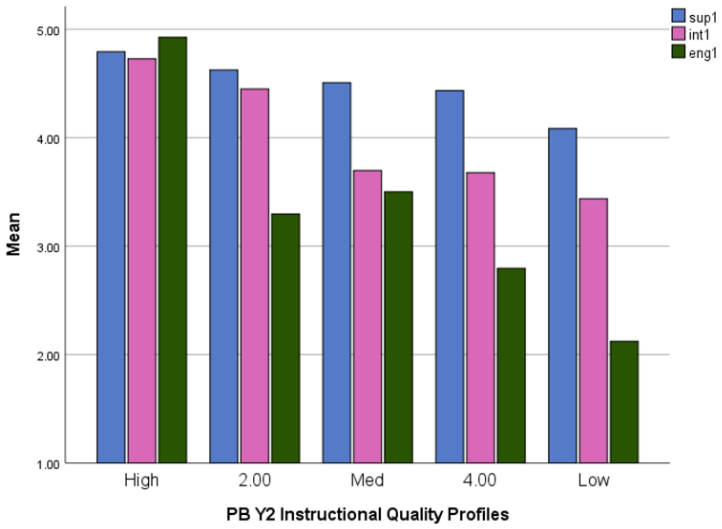
**Figure 1a. Year 1 Instructional Quality Profiles.**




Note: sup0 = Year 1 PQA "Supportive Environment" domain scores; int0 = Year 1 PQA "Support for Interaction" domain scores; and eng0 = Year 1 PQA "Support for Engagement" domain scores.

Establishing the extent and form of structural stability and change is a prerequisite for studying "individual-level" stability and change. One of the main reasons for this prerequisite is that the meaning of individual-level stability and change (or, in this case, program-level stability and change) depends largely on the extent and form of sample-level stability and change. For example, as illustrated below, if the shape of the multivariate profile labeled "high quality" at Time 1 (e.g., because that Time 1 shape appears to be the highest-quality profile shape at Time 1) does not match the shape of the profile labeled "high quality" at Time 2, then a program that moves from the Time 1 "high-quality" profile to the Time 2 "high-quality" profile may not be fairly characterized as evidencing program-level stability because the Time 2 "high-quality" profile may reflect substantially higher or lower quality than the Time 1 "high-quality" profile. Consequently, we begin by describing the structural stability and change of program profiles over time and then consider some of the more interesting program-level pathways of stability and change into and out of some of those sample-level profiles across their first three years of exposure to Prime Time services.

**Figure 1b. Year 2 Instructional Quality Profiles.**




In addition to the visual comparison of profile numbers and shapes across two or more points in time, assessing the structural stability and change of program profiles across time involves calculating and comparing profile shapes in terms of the profile *centroids*. A profile centroid is the geometric center, or mean, of the variable values corresponding to a profile shape. The distance between two profile centroids (i.e., the set of means on the profile variables within each profile group) can be described in terms of the average squared Euclidean distance (ASED). Similar to common variance statistics, ASED values can be used to represent the amount of variability in profile shapes among members of a profile group (e.g., we use these values as *homogeneity coefficients*, which reflect the extent to which the members of a profile group share the same profile shape). Using this ASED metric, we can generate quantitative estimates of the distances between profiles, both within and across time, and these ASED values reflect the extent of similarity between two profile shapes (where "0" indicates no distance between centroids, or identical profile shapes, and higher numbers indicate increasingly dissimilar profile shapes).

Table 1. ASEDs (Distances) between Year 1 and Year 2 PQA Profiles

| | Y2P1 | Y2P2 | Y2P3 | Y2P4 | Y2P5 |
|---|---|---|---|---|---|
| Y1P1 | 0.845 | 0.051 | 0.063 | 0.227 | 0.831 |
| Y1P2 | 2.177 | 0.284 | 0.382 | 0.076 | 0.225 |
| Y1P3 | 2.838 | 0.692 | 0.498 | 0.118 | 0.026 |
| Y1P4 | 3.777 | 1.213 | 0.862 | 0.329 | 0.042 |
| Y1P5 | 5.593 | 2.281 | 1.875 | 1.031 | 0.324 |

In order to quantify the extent of structural stability and change between the PQA profiles found at Year 1 (see Figure 1a) and the PQA profiles found at Year 2 (see Figure 1b), we calculated the ASEDs between each Year 1 PQA profile and each Year 2 PQA profile. As shown in Table 1, the ASEDs between Year 1 and Year 2 PQA profiles ranged from 0.026 to 5.593 and averaged 1.066. The results of a similar analysis of the ASEDs

describing the relations between each Year 2 PQA profile and each Year 3 PQA profile revealed ASED values ranging from 0.028 to 2.881 and averaging 0.703 (see Table 2).

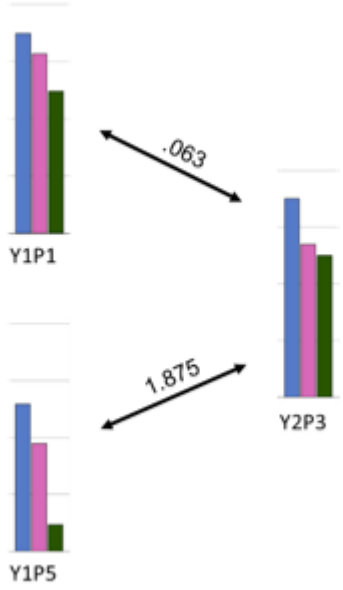
Table 2. ASEDs (Distances) between Year 2 and Year 3 PQA Profiles

| | Y3P1 | Y3P2 | Y3P3 | Y3P4 | Y3P5 |
|---|---|---|---|---|---|
| Y2P1 | 0.028 | 0.319 | 1.039 | 1.722 | 2.762 |
| Y2P2 | 0.648 | 0.160 | 0.103 | 0.149 | 0.601 |
| Y2P3 | 0.852 | 0.293 | 0.030 | 0.223 | 0.478 |
| Y2P4 | 1.585 | 0.709 | 0.148 | 0.053 | 0.085 |
| Y2P5 | 2.881 | 1.644 | 0.711 | 0.319 | 0.031 |

Given profiles composed of three PQA domain scores based on the five-point scaling of the PQA items (along with the use of unstandardized response scale values to derive and display the profiles), ASED values near 0.1 indicate good matches between profile shapes, and ASED values near 0.05 indicate excellent matches between profile shapes (cf. Bergman et al., 2003, Table 9.3). Conversely, ASED values near 0.3 indicate relatively poor matches between profile shapes, ASED values near 0.5 indicate poor matches between profile shapes, ASED values near 1.0 indicate very poor matches between profile shapes, and ASED values above 1.5 indicate extremely poor matches between profile shapes.

A summary and interpretation of the key ASEDs in Tables 1 and 2 is provided below. First, however, a clearer understanding of the correspondence between ASEDs and the difference between two profile shapes, particularly in relation to the PQA profiles under consideration here, can be obtained by examining Figure 1. For example, the shape of Year 1 Profile 1 appears visually to be quite similar, yet not identical, to Year 2 Profile 3, and this similarity is reflected by the 0.063 ASED value indicating a "good" match between profiles. In contrast, the shape of Year 1 Profile 5 appears visually to be quite different than Year 2 Profile 3, and this dissimilarity is reflected by the 1.875 ASED value indicating an "extremely poor" match between profiles.

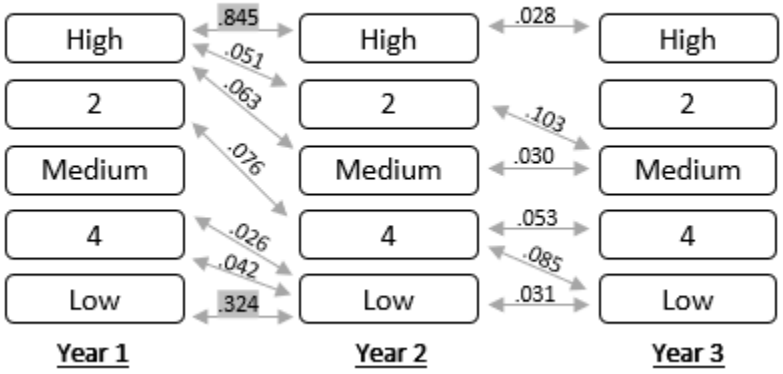
Figure 1: Understanding ASEDs by visually comparing profile shapes




A selection of the key ASEDs corresponding the relations between the Year 1 and Year 2, and Year 2 and Year 3, PQA profiles are shown in Figure 2. The 0.845 ASED between Year 1 Profile 1 and Year 2 Profile 1 indicates that the highest-quality Year 1 PQA profile is an extremely poor match to the highest-quality Year 2 PQA profile. The 0.051 ASED between Year 1 Profile 1 and Year 2 Profile 2, and the 0.063 ASED between Year 1 Profile 1 and Year 2 Profile 3, indicate that the highest-quality Year 1 PQA profile is an excellent match to Year 2 Profile 2 and a nearly excellent match to Year 2 Profile 3. If we classify the shape of Year 2 Profile 3 as the minimally-sufficient standard for high-quality instructional practices (e.g., the instructional total score [ITS] is 4.17, exceeding Prime Time's 4.1 high-quality ITS benchmark), we can describe the structural stability and change in PQA profiles largely by reference to that high-quality benchmark profile. For example, rather than referring to Year 1 Profile 1 as "high quality" and Year 2 Profile 3 as "medium quality" (e.g., as shown in Figures 1a, 1b, & 2), the ASED results showing a close match between the shapes of Year 1 Profile 1 and Year 2 Profile 3 indicate that we can refer to both of those profiles as reflecting high-quality instructional practices.

From this perspective, we can see that Year 2 Profile 1 reflects not only high-quality but extremely high-quality instructional practices. As described and applied in more detail below, this reclassification of sample-level profile shapes in terms of ASEDs allows to see clearly that programs moving from Year 1 Profile 1 to Year 2 Profile 1 do not reflect the stability of high-quality instructional practices but, rather, change from high-quality to extremely high-quality, or *exemplary*, instructional practices. Similarly, programs moving from Year 1 Profile 1 to Year 2 Profile 3 do not reflect change from high- to medium-quality instructional practices but, rather, stability of high-quality instructional practices.

Figure 2: Structural Stability and Change of PQA Profiles

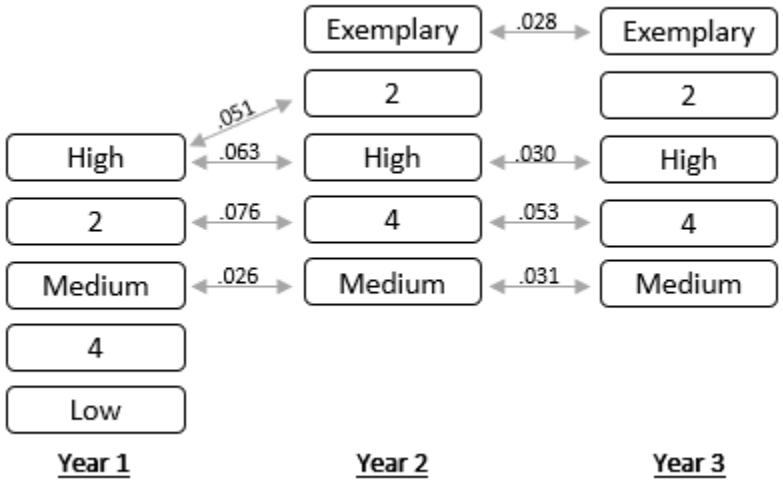



Applying the same line of reasoning to the definition and classification of a "low-quality" profile shape reveals a similar shift in meaning across time. For example, the 0.324 ASED describing the relation between Year 1 Profile 5 and Year 2 Profile 5 indicates a relatively poor match between those two profile shapes, and the 0.026 ASED between Year 1 Profile 3 and Year 2 Profile 5 indicates an excellent match between those two profile shapes. This pattern of structural stability and change of the "low-quality" profile indicates a shift in meaning across time in what appears to be low quality. If we use ITS scores below 3.0 as the benchmark for low-quality (cf. Smith et al., 2013), then the ITS of 3.22 for the Year 1 Profile 5 "low-quality" profile suggests that programs characterized by this profile shape barely qualify as being considered low quality (e.g., we have seen many programs from other systems characterized by significantly lower-quality instructional practice profiles that reflected by Year 1 Profile 5). Nevertheless, given that the Year 1 Profile 5 shape appears to reflect the lowest-quality instructional practices among all programs across all three years of exposure to continuous improvement processes considered here, we use Year 1 Profile 5 as the standard for considering patterns of structural stability and change in low quality across time.

Using this standard for low quality (i.e., Year 1 Profile 5), the relatively poor match (i.e., the 0.324 ASED) between Year 1 Profile 5 and the lowest-quality profile shape at Year 2 (i.e., Year 2 Profile 5) indicates that few if any low-quality programs remain in the sample after one year of exposure to Prime Time's continuous improvement processes. This conclusion is consistent with the 3.66 Year 2 Profile 5 ITS score as well as with the 0.026 and 0.042 ASEDs between Year 2 Profile 5 and Year 1 Profiles 3 and 4, respectively, indicating that Year 2 Profile 5 corresponds most closely with what appeared during Year 1 to be medium-quality instructional practices (see Figure 3).

Applying these standards for high and low quality to all profiles across all three years of data considered here provides the basis for the following conclusions: (a) in Year 1, 0% of the programs were characterized by extremely high-quality instructional practices, 16% were characterized by high-quality instructional practices, and 14% were characterized by low-quality instructional practices; (b) in Year 2, 10% of the programs were characterized by extremely high-quality instructional practices, an additional 36% were characterized by high-quality instructional practices, and 0% were characterized by low-quality instructional practices; and (c) by Year 3, 15% of the programs were characterized by extremely high-quality instructional practices, an additional 31% were characterized by high-quality instructional practices, and 0% were characterized by low-quality instructional practices. In other words, whereas only 16% of the Year 1 programs were characterized by high quality instructional practices, approximately 50% of the programs had achieved the high-quality instructional practices benchmark by Years 2 and 3. In short, at the sample level, program quality (as defined by PQA profiles) increased substantially between Years 1 and 2 and remained relatively stable, although increasing slightly, between Years 3 and 4. As shown in Figure 3, realigning and relabeling the profiles across years according to their ASEDs more clearly illustrates this shift to higher-quality instructional practices across time.

Figure 3: Revised Structural Stability and Change of PQA Profiles




## Program-level (aka, Individual-level) Stability and Change in PQA Profiles

Given the substantial structural changes in PQA profiles during the first two years of exposure to Prime Time's continuous improvement process (e.g., the emergence of the exemplary high-quality profile and the disappearance of the low-quality profile), the description of program-level stability and change needs to be anchored to these structural changes. For example, programs characterized by Profile 3 in Year 1 and Profile 3 in Year 2 that might initially be viewed as following a stable "medium-quality" *instructional practice pathway* can be seen, in light of the sample-level structural changes, as following pathways from medium-quality instructional practices to high-quality instructional pathways. Consequently, using the patterns of structural stability and change as an anchoring framework, we classified each Year 1 program according to its pattern of program-level stability or change. Given that we found substantial structural changes between Year 1 and Year 2 and substantial structural stability between Year 2 and Year 3, we created one set of program-level instructional practice pathway variables focused on the transition from Year 1 to Year 2 and another set of program-level instructional practice pathway variables focused on the transition from Year 1 to Year 2 or Year 3 combined. Given the percentages of programs with unavailable PQA data for Year 2 (14%) and Year 3 (25%), this latter set of variables allows us to track program changes for a higher percentage of programs.

There are a wide range of specific program-level pathway variables that can be constructed from the combination sample-level structural stability and change information and information about the individual program-level movements into and out of year-specific profiles of instructional practices. For example, of the 40% of Year 1 programs considered to be of low- or medium-low quality instructional practices (i.e., Year 1 Profiles 4 and 5), 50% of these programs followed a stable low to medium-low quality instructional practice

pathway through Years 2 and 3, but the other 50% followed pathways to higher-quality through Years 2 and 3. We created a dichotomous "High-Risk PQA Prodigal Pathways" variable to reflect these divergent pathways, and will use this variable to examine both potential causes of such divergence (e.g., higher vs. lower implementation fidelity) and potential consequences of such divergence (e.g., higher vs. lower staff satisfaction and retention).

In addition, examining program-level stability and change across all programs and all years revealed that 72% of programs evidenced program-level changes from lower- to higher-quality profiles of instructional practices, 20% of programs evidenced program-level stability in profiles of instructional practices, and only 8% of programs evidenced program-level changes from higher- to lower-quality profiles of instructional practices. These results indicate impressive increases in the quality of instructional practices by the second year of exposure to Prime Time's continuous improvement processes, and these increases were largely maintained or increased further by the third year of exposure to Prime Time's continuous improvement processes.