David P. Weikart
Center for Youth
Program Quality

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

Charles Smith, Ph.D.
Executive Director, David P. Weikart Center for Youth Program Quality
Senior Vice President for Research, the Forum for Youth Investment

Updated December 2013

This paper was made possible by the financial support from the Raikes Foundation.

The first in a series of technical working papers developed by the Weikart Center as part of the Expanded Learning Initiative.

DAVID P. WEIKART
CENTER FOR YOUTH
PROGRAM QUALITY

Technical Working Paper No. 1:

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

Charles Smith, Ph.D.
Executive Director, David P. Weikart Center for Youth Program Quality
Senior Vice President for Research, the Forum for Youth Investment
Contact: charles@cypq.org

This paper was made possible by the financial support from the Raikes Foundation.

Gina McGovern and Anna Gersh at the Weikart Center contributed editorial and graphic design support for this report. Neil Naftzger at American Institutes for Research contributed the Rasch Analyses and the supporting Technical Appendix.

Other Technical Working Papers developed by the Weikart Center as part of the Expanded Learning Initiative:

- No. 2 - Position, Performance, Proof: Next Stage Technical Guidance for Defining and Measuring Youth Skills
- No. 3 - The Individual-Growth-by-Levels-of-Quality Evaluation Design for Expanded Learning Systems and Settings
- No. 4 - Measuring Youth Skills in Expanded Learning Systems: Case Study for Reliability and Validity of Youth Development Executives of King County Skill Measures and Technical Guidance for Local Evaluators

Summary

This paper introduces the nomenclature of performance-based accountability systems (PBAS) to the expanded learning field, provides a policy case study for a countywide system in southern Florida and uses data from that system to explore the issue of quality thresholds. We present an expanded design standard to guide development and improvement of PBAS policies and further develop a theory of lower-stakes accountability to guide effective use of incentives of various types. Findings suggest that (1) the PBAS framework defines critical concepts and improves our ability to describe existing quality improvement systems, (2) the Youth Program Quality Assessment (Youth PQA) can be used to produce a program rating of sufficient reliability for use in a PBAS, and (3) that the Palm Beach County PBAS design is an exemplar for expanded learning policies.

General recommendations for PBAS designs include:

- PBAS design should differentiate roles and link performance measures to incentives targeted at specific management and service delivery roles.
- PBAS designs should include program ratings for multiple service domains linked to a mix of higher- and lower-stakes incentives.
- PBAS should emphasize participants' understanding of performance levels and sense of fairness while evolving toward higher-stakes incentives over time.

Detailed recommendations for Weikart Center clients using the Youth Program Quality Intervention and related Program Quality Assessments as the basis for an expanded learning PBAS design include:

- Recommendations for best practice in each element of the seven elements in PBAS design standard.
- Detailed description of a composition map for program ratings and performance levels for nine commonly used measures in expanded learning PBAS.
- A PBAS design exemplar based on the Palm Beach County case describing specific combinations four types of incentives (financial, customer review, supervisory review, access to data) with two types of performance levels (high and low) and nine program ratings to achieve an optimal, lower-stakes, PBAS design with higher-stakes elements.

Introduction

Over the past decade, states and communities have built expanded learning[1] systems to increase access to high-quality, out-of-school-time experiences, improve 21st century skill and school success outcomes[2], close gaps in academic achievement, and provide supervision for the children of working parents. From a social policy perspective, two characteristics of expanded learning systems stand out. First, expanded learning systems are a missing piece in local, social and human capital infrastructures that focus on middle childhood and adolescence, much like early childhood education was a missing piece in prior decades. Expanded learning systems represent opportunities for communities to strategically pursue shared goals for learning, socialization, and social participation by children, young adults, and many early career professionals, while simultaneously supporting parents who work. Second, expanded learning settings are defined by their organizational diversity and flexibility of program content and staffing, both of which facilitate unique responsiveness to community needs and access to community and cultural resources. In most communities, provider organizations are widely varied (e.g., 21st Century Community Learning Centers, Boys and Girls Clubs, local community-based organizations, churches, sports leagues, etc.); include a mix of fee-based, means-tested and free services; have great flexibility to deliver many different kinds of content; and often employ an early career, adult staff with qualifications and experience from a variety of sources. The opportunity for communities to pursue shared goals for youth, while simultaneously promoting diverse but coordinated pathways to achieve these goals with accountability is a leading-edge model for both public policy and for intervention science.[3]

However, the two defining characteristics of expanded learning systems mentioned above – community strategy with shared goals on one side and organizational diversity and responsiveness on the other – might also seem to work at cross purposes. How can communities synchronize the goals, funding, and requisite accountabilities for that funding and do so across the diverse community organizations, to achieve population-level effects? One critical integrative policy innovation for expanded learning policies – facilitating the advancement of place-based strategies through a diverse community organizational

[1] We use the term *expanded learning* to describe any setting where a group(s) of children/youth and at least one consistent adult participate over multiple sessions for a learning purpose. We use the term to refer to settings variously labeled out-of-school time, after school, extracurricular clubs, summer camps and sports; some mentoring, tutoring, and apprenticeship models; and programming for disconnected and homeless youth.

[2] Hereafter the term *21st century skills* is used to describe a range of cognitive, intra-personal and inter-personal knowledge skills, and beliefs. Our use of the term encompasses other terms: social and emotional skills, soft skills, intermediate skills, after school outcomes, etc. Our use of the term *school success outcomes* refers to grades, tests scores, and school behavior.

[3] Pursuit of shared community goals through different institutional pathways, with accountability, is the essence of the collective impact approach (Kania & Kramer, 2011). Dodge (2011) uses a different frame to describe similar challenges where intervention science and child and youth policy intersect, as does National Research Council and Institutes of National Research Council and Institute of Medicine (2009).

architecture – has been the development of quality improvement systems (QIS). This is true for several reasons:

- QIS provide communities with normative frameworks for positive youth development experiences,[4] and they articulate standards for management practices, service quality, and program effectiveness that a wide variety of expanded learning providers can agree on and are willing to be accountable for (Yohalem, Devaney, Smith, & Wilson-Ahlstrom, 2012).

- QIS frequently create opportunities for cross-age, cross-silo, cross-sector, and cross-town planning and coordinated action, effectively blending resources from multiple public and private funders through the shared purposes of accountability and improvement (Yohalem, Ravindath, et al., 2010).

- QIS typically include quality intermediary organizations (QIO) as dissemination agents for program quality improvement interventions,[5] as well as technical supports necessary for program managers to participate in the QIS. QIO also often provide services related to program quality assessment, participation tracking, curriculum, and professional development.

In short, quality improvement has become a leading place-based strategy enacted by diverse groups of organizations in many communities and states.[6] Quality improvement is often one of the primary domains of collaboration that many of the diverse actors involved in expanded learning systems can agree on, and therefore both QIS and QIO have grown in number and capacity in recent years. These local quality improvement policies, and more importantly the value of the services they affect, represent a large social investment in the well-being of youth in middle childhood and adolescence.[7]

---

[4] The term *positive youth development experiences* refer to broad consensus about how to advance skill development during middle childhood and adolescence (c.f., Eccles & Gootman, 2002; Wilson-Ahlstrom, Yohalem, DuBois, & Ji, 2011).

[5] The term intervention could be replaced with the term innovation. QISs require implementation of good management practices, and these management interventions are also organizational innovations. QIOs have a track record of successful dissemination of management innovations (Honig, 2004; Knockaert & Spithoven, 2012).

[6] For example, 85 networks in 38 states and Canada implement the *Youth Program Quality Intervention* and numerous other systems use similar tools in the expanded learning field. Federally and locally funded Quality Rating and Improvement Systems for early childhood and school-age settings are present in most states. Together, these efforts certainly represent tens of thousands of program sites located in most communities.

[7] As a "back of napkin" estimate using the Weikart Center's client base as an example, with the following assumptions: 3000 EL sites, total QIS cost per year $6,000 per site, total budget per site $150,000 per year. Then total investment in quality improvement and accountability for 3,000 sites totals $18 million annually as an investment for $450 million in services each year. (About 4 percent of total spending for quality improvement and accountability.)

As investments have grown and systems have evolved, however, the goals for expanded learning settings have changed. For the past decade, funder and community goals have focused on defining and setting standards for high-quality services (Eccles & Gootman, 2002; Gambone, Klem, & Connel, 2002) and building local capacity - the QISs and QIOs – to implement high-quality, expanded learning strategies at scale. More recently, funders and stakeholders are emphasizing accountability for service quality and evidence of effectiveness that the expanded learning service is actually improving 21st century skills and school success outcomes. For the expanded learning field, the question has evolved from "What is quality?" to "What level of quality is sufficient to achieve goals for 21st century skill building and school success?" and "How can funded programs best be held accountable for high-quality services?" This evolution of priorities and questions requires expanded learning systems to respond with more highly elaborated QIS policies.

In this Technical Working Paper, we attempt to address a number of technical issues related to how QIS policies can be updated to target specific levels of performance. In the first section of this paper, we attempt to reframe the emerging QIS policies in the expanded learning field and their core components as performance-based accountability systems (PBAS), drawing on recent work from the Rand Corp. (Camm & Stecher, 2010; Stecher et al., 2010). The body of QIS policies for the expanded learning sector have received early treatments (Smith, Akiva, Devaney, & Sugar, 2009; Yohalem et al., 2012), but the field lacks a well-specified nomenclature and framework for describing and advancing the work. We extend the earlier PBAS framework as a design standard to guide development of PBAS in the expanded learning and other fields.

In the second and third sections of the paper, we apply these concepts to the nation's most mature performance-based accountability system (PBAS) for an expanded learning system, located in Palm Beach County, Fla. We apply the PBAS framework to the Palm Beach County case (with four additional case examples in Appendix A) and then attempt to answer a number of related technical questions using data from the Palm Beach County system. We utilize a five-year longitudinal data set from the Palm Beach County PBAS to describe performance and the reliability of performance ratings, and as a preliminary source of validation evidence for the Palm Beach County PBAS design.

Finally, we present an integrative discussion of the prior sections with general recommendations focused on defining service quality, application of measures and composition of site-level ratings. While these findings are particularly germane to systems using the Youth Program Quality Intervention and Youth Program Quality Assessment suite of tools and methods[8] in the expanded learning field, we think

[8] The *Youth Program Quality Intervention* (YPQI) and *Youth and School-Age Program Quality Assessments* and extensions (Youth PQA; School-Age PQA; Academic Climate, STEM, Arts, Health and Wellness) are the most widely used quality improvement intervention and metrics in the expanded learning field.

these findings also hold more general applicability across different performance improvement approaches and across service domains in the education and human services fields. Following these general recommendations, we provide detailed recommendations regarding application of the design standard for expanded learning systems.

While this technical white paper is targeted at a limited audience working in the domains of performance measurement and performance management, the broader purpose of this work should not be lost: Measuring the performance of expanded learning settings and systems is a strategic path toward wider understanding of, and appreciation for, the unique services that expanded learning settings provide.

Part I. The PBAS Framework in the Expanded Learning Field

In this section, we lay out several critical concepts from the PBAS framework developed by Camm and Stecher, et al., then highlight some useful extensions from the framework that integrate existing strengths of the QIS in the expanded learning field. Hereafter we replace the acronym QIS with PBAS.

Important Concepts From the PBAS Framework

In this section, we define PBAS and their core components – goals, measures, and incentives – and highlight some additional key insights and concepts from the PBAS framework. A performance-based accountability system is defined by Camm and Stecher as:

> … a mechanism designed to improve performance by inducing individuals or organizations that it oversees to change their behavior in ways that will improve policy outcomes about which the creators of the PBAS care. To do this, the PBAS (1) defines specifically whose behavior (individuals or groups of individuals in an organization) it wants to change (2) tailors an incentive structure to encourage these individuals or organizations to change their behavior, and (3) defines a set of performance measures it can use within the incentive structure to determine whether changes in behavior are promoting the PBAS' goals (Camm & Stecher, 2010, p. ix).

Three key components of PBAS models are goals, measures, and incentives and are defined this way in the PBAS framework:

> First, policymakers must agree on a set of *goals* or desired long-term outcomes for the service-delivery activity; these are usually expressed in general, nonquantified terms (e.g., world-class achievement, efficient public transportation, high-quality child care). These goals define what the service-delivery activity is supposed to achieve under the new regime of the PBAS.
>
> The second piece of the PBAS is an *incentive* structure that assigns rewards or sanctions (or some combination thereof) to individuals or organizations to try to motivate changes in their behavior. The incentives need not be financial; we include in our definition nonfinancial consequences that might motivate changes in provider behavior, such as greater autonomy, loss of control, or public reporting.

> The third element of the PBAS is a set of *measures* that can be used as the basis for applying the incentives to the people and the units that deliver the services. The designers of a PBAS must choose a way to define performance in order to implement the incentive structure and encourage better performance on the part of service providers (Stecher et al., 2010, p. 5).

Several additional insights and concepts from the PBAS framework are particularly important. First, the PBAS framework defines generic terms for a service production model with inputs, outputs, and outcomes:

> If designed and implemented appropriately, the PBAS encourages those delivering the service to take actions that improve measured outputs in the short term and promote desired outcomes in the long term. Evaluation provides evidence about the effectiveness of the PBAS. Standard approaches to program evaluation envision program processes that transform inputs into outputs, which, in turn, ultimately affect the program outcomes that really interest policymakers. Our application of this approach accepts that the *outcomes* of a program can almost never be directly observed and measured. As a result, any accountability system must rely on measures of program *outputs* that can be measured. Throughout, we use this distinction to draw a line between *outputs* and *outcomes* when outcomes relevant to policymakers cannot be directly observed and measured (Stecher et al., 2010, p. 7 & note 5).

By definition, the PBAS is limited to measuring inputs and outputs that can be observed. In their more detailed case studies, the No Child Left Behind policy for K-12 education is defined as a PBAS with outcome goals for adult life success that flow from both unmeasured outputs (quality of instruction) and the central measured output, student achievement. In contrast, the PBAS for early childhood education (state Quality Rating and Improvement Systems (QRIS)), define school readiness as the broad outcome goal, while quality of care is the focal output that is measured. So, importantly, service outputs can include both core processes like instruction, as well as proximal skills like academic achievement.[9] Note also that the evaluation of expanded learning systems is defined as serving the specific function of determining effectiveness of the services (and PBAS) in the production of ultimately desired outcomes.

Several additional attributes of the PBAS framework are evident in the explicit focus on individual behavior. First, targeting individual behavior change requires clear specification of which actors and which performances are important to the production of service outputs. Further, an effective PBAS may target multiple roles and behaviors at the same time, such as management practices and front-line service delivery. Second, effectively targeting individual behavior may reduce the need for rewards or consequences that are monetized or reputational. If the measures deployed by the PBAS produce opportunities that are valuable to participants – data that supports professional skill development such as

[9] In the author's conversations with ISO 9000 auditors working with community colleges in Mexico, this issue was routine: The outputs of schooling processes include both instruction and proximal student learning but the quality assurance process may differ in important ways, depending on which output is measured.

individual performance feedback, jobs with greater autonomy, or pathways to demonstrate accountability – that were not previously available, then implementation of the PBAS measures may already incentivize individual performance change. Third, if the PBAS produces information about performance at a fine-grained level (by, for example, differentiating performance measurement between management processes and instructional processes) the PBAS becomes a part of the management system that produces the service.[10]

Integrative Extension of the PBAS Framework as a Design Standard for the Expanded Learning Field

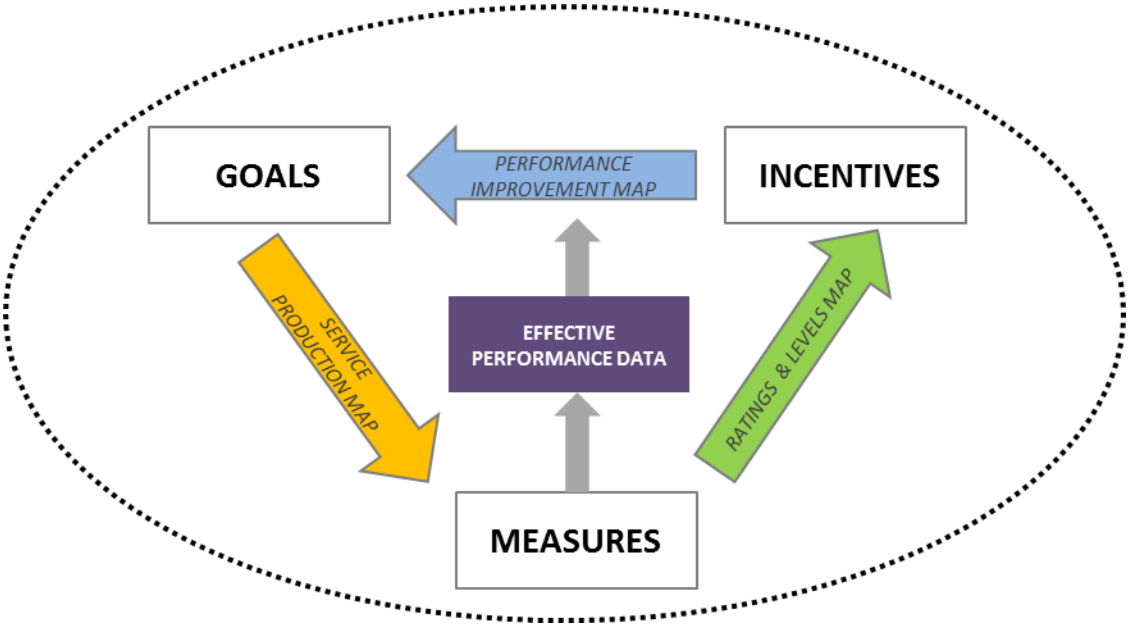
In this section, we draw upon, but also extend beyond, the PBAS framework, defining PBAS components in terms of the expanded learning field. In some cases we bring peripheral elements of the Camm and Stecher framework to the center; in others, we extend into new territory. We use Figure 1 to set terms and concepts that are used in the remainder of the paper.

Figure 1 represents the three PBAS components – goals, incentives, and measures – together with a set of process maps (represented with arrows) that define a design standard for PBAS in the expanded learning field. The sequence of components starts with goals and proceeds to measures and then to incentives, with critical process mapping steps between each component. At the center lies effective performance data, an element of the system that results from well-selected performance measures.

Next, we define each element of Figure 1 in terms of the expanded learning field, in order of appearance.

[10] Two important points follow: First, the culture of the oversight agency may change to become more focused on *coaching* for continuous improvement of outputs and less focused on *monitoring* inputs. This culture change has been reported in both Palm Beach County and Department of Child and Youth Development in New York City (Yohalem et al., 2012). The second point is that the external relationship with clear boundaries is no longer with the funding agency but with the firm that conducts the performance measurement, as with "third party" assessment in private sector quality assurance models like ISO 9000. PBAS facilitate more integrated "supply chain" relationships by spreading interests in producing high quality outputs across the chain of producers and consumers.

Figure 1. PBAS Design Standard for the Expanded Learning Field



Goals. The PBAS framework describes these as ultimately desired goals that are stated broadly, and that typically are assessed not by the PBAS but through rigorous program evaluation. A key purpose for broadly stated public goals is to convey the important social purposes that are served by the organization and to direct the energies of staff toward desired outcomes. For the expanded learning field, goals are typically stated as outcomes that occur in other settings: greater success in school, self-regulation in the family, better decisions in the neighborhood, readiness in the workforce, etc. However, quality is also a goal for child care policies, so in the expanded learning field-service quality is also often a broadly stated public goal.

Service Production Map. The service production map is a specific type of logic model that describes a sequence of measurable outputs, organizational processes and individual skills, all of which link program inputs to service outcome goals. A key purpose for the service production map is to identify the active ingredients of program settings and individual skill domains so that output measures provide effective performance data. (See Appendix B for further discussion of the attributes of effective performance data.)

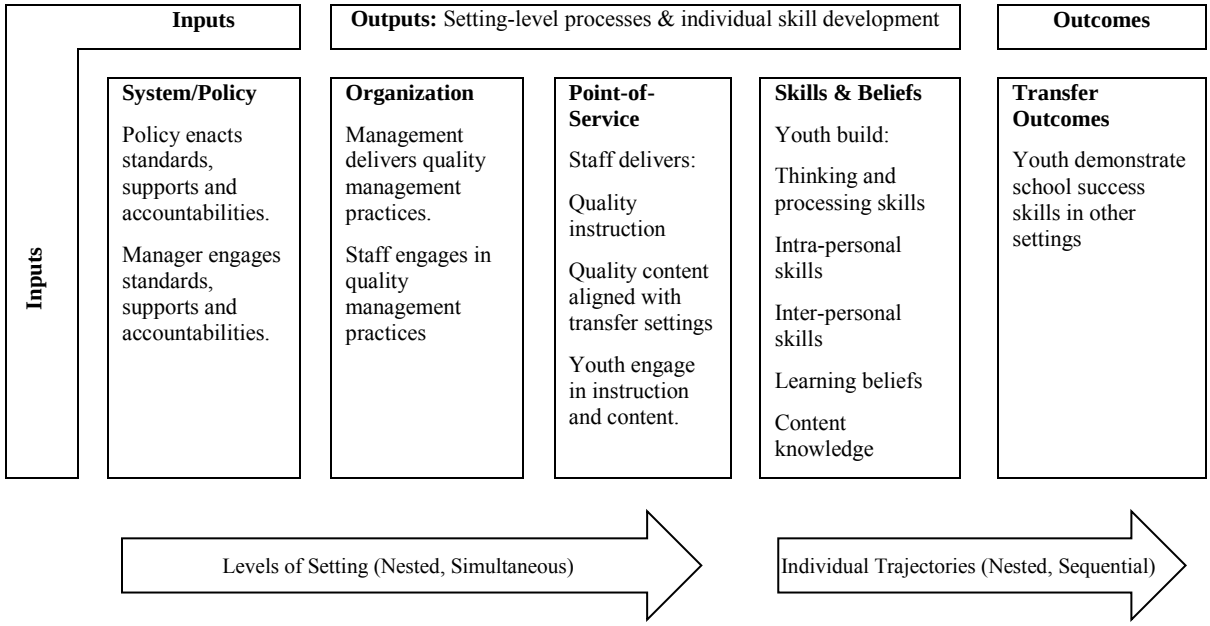
Over the past decade, the Weikart Center has developed two generic service production maps specifically for PBAS applications in the expanded learning field. The first describes nested levels of settings – system level, organization level, and point-of-service level – and was the organizing map for the *Youth Program Quality Intervention Study* (Smith, Akiva, Sugar, et al., 2012). The second describes an individual developmental trajectory for youth that links the youth experience to youth demonstration of

individual skills in external settings (Smith, Hallman, et al., 2012). Additional detail about this second map is provided in Appendix C.

Figure 2 combines these service production maps into a single template to capture major elements of the production process for expanded learning services. While there are many things to say about Figure 2, the important points for this discussion are:

- The purpose of this map is to identify a few measureable elements from what is otherwise a very complex dynamic system. It is not a complete map of any specific theory or design.
- By definition:
    - Inputs include "structural features" (e.g., teacher education) and are not directly related to outputs at the point-of-service level, or in a direct relationship with individual effects.
    - Outputs include both processes occurring at the system, organization and point-of-service levels, as well as proximal skill development in individuals.
    - Outcomes are not measured by the PBAS and are described as occurring in external or "transfer" settings.
- The first three columns of Figure 2 describing the system, organization, and point-of-service levels of setting are locations for simultaneously occurring processes (e.g., quality management practices and quality instructional practices), which together directly produce youth experience in the point-of-service level. These setting-level processes can be measured at a single point in time, and point-in-time association between setting processes is of interest.
- At each level of setting, leadership roles *enact* processes and participant roles *engage* those processes. Processes that are enacted by a specific role are good targets for accountability in a PBAS.
- The last two boxes of Figure 2 describe youth skill development and skill transfer which are produced sequentially and, in contrast to point-in-time measures of setting processes, are best understood as individual growth trajectories occurring through time, requiring measurement at multiple time points for the same individual.

Figure 2. Logic Model for Production of EL Services



Measures. Performance measures in a PBAS are selected in order to produce information that can be used to change the behavior of individuals producing the service. Performance measures are selected to align with key service outputs, organizational processes, and individual skills, identified in the service production map. Measures are most effective when they can be composed as information about performance levels which can be further aligned with various types of incentives. This subject is a primary topic of this paper and is taken up in Sections II-IV.

Rating and Levels Map. The purpose of a rating map is to describe rules for the composition of measures and for the setting of performance levels. This map makes clear how program ratings are constructed from various measures, including information about the methods for data collection and the reliability of each component measure. The purpose for selecting performance levels is to provide performance goals and clear expectations about how incentives are applied. This subject is a primary topic of this paper and is taken up in the sections II-IV.

Incentives. In the PBAS framework, an incentive is the reward or sanction linked to satisfaction of a specific performance level. Incentives are intended to motivate individual behavior change.[11] Incentives can range widely including those that are monetized (reimbursements, publicity of service quality) and those that have other kinds of value to individuals such as a heightened clarity about individual performance or access to a learning community. Incentives like supervisory review can be mundane. In the expanded learning field, fewer systems have mapped performance measures to specific

[11] A good parallel discussion of PBAS incentives is available in Zellman, Perlman, Le, and Setodji (2008)

incentives defined in terms of substantial monetary value or publicity, with the important exception of a few states that have included school-aged child care settings in a QRIS. See descriptions of Vermont and Arkansas in Appendix A. Extensive voluntary participation in expanded learning PBAS suggests the effectiveness of non-monetary incentives related to adult learning, professional skill development, and improved performance.

Performance Improvement Map. Performance improvement methods link performance measures to improved individual performance, making it possible for individuals to experience accountability as attainable. One of the unique strengths of expanded learning PBAS is the prevalence of evidence-based improvement methods that differentiate the application of the PBAS by two roles: quality management practices of managers and quality instructional practices of teachers and youth workers. In the field's leading improvement intervention,[12] quality management practices are defined as management leadership of a staff team through a sequence of performance assessment, data-driven improvement planning, instructional coaching and performance feedback for individual teaching staff, and training in specific instructional practices. Quality instructional practices differ by program model but considerable consensus exists in the field for a subset of active learning practices for middle childhood and adolescence (Yohalem, Wilson-Ahlstrom, Fischer, & Shinn, 2009).

Effective Performance Data. Effective performance data could be part of the measures component but is highlighted here to point out that not all data is equal. PBAS data should support not only judgments about performance at a high level of aggregation, but also, and perhaps most importantly, at the level of individual behavior and skills where performance improvement actually occurs. This means that the data generated by the PBAS should have certain characteristics such as method-feasibility, timeliness, description of objective conditions and behaviors, reliability, sensitivity, and validity. Where possible the data should also describe performance at multiple levels of setting. These characteristics of effective performance data are further defined in Appendix B.

Lower Stakes Accountability

The Weikart Center has advocated for lower stakes models for accountability in expanded learning PBAS over the past decade (Smith & Akiva, 2008; Smith & Hohmann, 2005). Additional detail on the lower stakes concepts and theory is provided in Appendix D. Our use of the term "lower stakes" refers to an individual's experience of an "accountability." Our use of the term "accountability" as a

[12] The *Youth Program Quality Intervention* is the most widely used set of performance improvement supports and methods in the expanded learning field and has been subjected to evaluation in a randomized trial, demonstrating effects of quality management practices on quality instructional practices (Smith, Akiva, Sugar, et al., 2012).

singular noun[13] refers to the combination of a performance level and an incentive for a specific aspect of service production. For example, a program manager might have numerous accountabilities for successful delivery of expanded learning services, including keeping the program well-attended. The performance level for attendance might be serving 50 unique youth each week and the incentive might be absence of supervisory review if the performance level is met. We would also refer to the incentive (in this case, supervisory review) as lower stakes because the Program Manager's experience of the accountability does not include immediate threat to their professional status or income.

PBAS designs in expanded learning systems are lower stakes to the extent that many of the individuals experiencing those accountabilities:

- Have access to evidence-based professional supports known to improve performance;
- Can attain the performance level with reasonable time and effort;
- Consider attainment of the performance level to be a "good use"[14] of time and effort;
- Believe that the measure of performance is precise (reliable) and fair (valid);
- Are not threatened with loss of professional status or income as a result of a low score, and without recourse (e.g., immediate loss of funding, publication of low performance levels).

For the expanded learning PBAS using the Youth Program Quality Intervention and Youth or School-Age Program Quality Assessments, our suggested methods for a lower stakes PBAS design have included:

- Targeting primary incentives/consequences toward implementation of quality management practices (e.g., YPQI) by program managers, because these can be achieved by most program managers and can be measured with precision;
- Assuring access to improvement supports and methods so that participants can have a fair shot[15] at attaining high performance levels for both quality management practices and quality instructional practices; and

[13] This use of the term "accountability" parallels use of the term "requirement" in private sector quality control schemes where the term is defined as "…a capability to which a project outcome (product or service) should conform." Accountabilities could be seen as the "requirements" for the expanded learning service, but the usage seems awkward.

[14] Clearly this term is subjective but it strikes at the motivational content we are interested in. The lower stakes accountability theory suggests that lower stakes accountabilities can motivate behavior change by simply identifying the performance level and then providing feedback and improvement supports.

[15] Stecher et al. (2010) emphasize the perception of fairness among participants in a PBAS. If participants do not perceive the rules as fair it increases the likelihood of perverse effects like gaming the system.

- Limiting publicity of low scores for individuals by averaging across multiple individual assessments to create a program rating, and for program managers by sharing program ratings only within program teams, immediate supervisors, and funders.

While a lower stakes theory is not fully elaborated,[16] it is helpful to note that the lower stakes criteria described here implies that certain types of incentives, especially the improvement supports like training and technical assistance or coaching, go to the lower performers. This logic reverses the typical understanding of accountability as a method for sanctioning weak performers. Lower stakes designs are particularly well-suited to the expanded learning field because these designs are oriented to build capacity across all participants rather than eliminate lower performers.

As noted by Camm and Stecher (2010), PBAS evolve over time to address specific policy problems. We argued in the introduction that the expanded learning field was undergoing a stage of what policy analysts call "problem redefinition" by moving away from the issues of defining and implementing quality services at scale and toward issues of defining levels of effective performance and using accountability tools to reach those levels at scale. An important implication of this evolution is that expanded learning PBAS are evolving toward higher-stakes designs where the emphasis shifts in part toward rewarding high performers. Indeed, the case study we develop in Part II reflects a system integrating higher-stakes elements into an overall lower stakes accountability design.

Part II. The Palm Beach County Policy Case: Goals, Measures, Incentives, and Link Mechanisms

In this section we apply the PBAS design standard defined in the prior section to a specific case, the Palm Beach County PBAS, which is one of the most mature systems of its kind. Again, for simplicity, we refer to the Palm Beach County Quality Improvement System (QIS) as a performance-based accountability system (PBAS) in the remainder of this paper.

Palm Beach County PBAS Background

The Palm Beach County PBAS has been in place since 2006. In 2013, 121 program sites participate; a total of 130 unique programs in the county have participated during the past seven years of PBAS operation. A concise description of the evolution of the Palm Beach County PBAS is provided in Smith, Akiva, Blazevski, Pelle, & Devaney, 2008 and several evaluation reports produced by an external

[16] There are few new ideas. We are still learning the organizational studies literature and remain confident that others have given these issues fuller treatment elsewhere.

evaluator at Chapin Hall (Spielberger & Lockaby, 2006, 2008; Spielberger, Lockaby, Mayers, & Guterman, 2009). They are available at the Prime Time Inc. website.[17]

Mapping Goals, Measures and Incentives for the Palm Beach County PBAS

Table 1 describes the Palm Beach County PBAS in terms of the Figure 1 design standard by specifying goals, measures, and incentives, and by filling in some of detail for the ratings and levels map, which links measures to incentives. The broad goals of the Palm Beach County QIS are improved developmental outcomes for children in the county, especially defined in terms of school success. Increased quality of care for children and development of a skilled child and youth services workforce are also important publicly stated goals. The primary PBAS measures are aligned to setting level processes (outputs) and include measures of the quality management process (fidelity to YPQI)[18] and quality instructional practices (Palm Beach County Program Quality Assessment (PBC-PQA), Form A). Primary incentives include funding, access to QIS supports, development of individual skills, and public recognition.

[17] http://www.primetimepbc.org/our-work/prime-time-impact. Given the dearth of evaluations conducted on PBAS, these reports are underutilized.

[18]The PBC-PQA Form B is also completed on an occasional basis as part of the coaching process but is not identified as a PBAS measure and is not mapped to incentives so it is not discussed further in this paper.

Table 1: Key Components of Palm Beach County PBAS

Goals	Measures	Incentives
Quality management practices Quality care as defined by Palm Beach County standards Child/youth development and learning • Engagement • 21st century skills • School success	Readiness checklist for PBAS entry Count of quality management practices (YPQI) Program Rating for quality instructional practices (PBC-PQA) Youth engagement survey Organizational stability • Manager and staff retention • No location change or service disruption)	Funding • Per child reimbursement • Staff education stipend Access to performance supports and methods • QIS services • ELO vendors Public recognition for high quality Individual skills training: management and instruction

Definition of levels	Levels mapped to incentives	
Baseline – All programs in first improvement cycle; organization instability; non-compliance with YPQI **Intermediate** – All programs following first cycle who comply with YPQI but not in maintenance **Maintenance** – All programs compliant with YPQI and with a PBC-PQA total score over 4.1 in two successive years.	PBC-YPQI Fidelity →	Per child reimbursement;[19] education stipend; management skills
	PBC-PQA > 4.1 →	Public recognition

The PBAS has three defined performance levels: baseline, intermediate, and maintenance. The *baseline* level is defined by a count of quality management practices in year one, and/or experience of a major organizational instability in subsequent years which, regardless of the system's achieved performance level, effectively returns the system to baseline status. Quality management practice entails the following requirements for a program manager: (1) conduct a self-assessment in the first program cycle, (2) write a program improvement plan, (3) provide quarterly summaries of progress toward goals, (4) meet quarterly with QIS coaches, and (5) conduct quarterly observation-reflection meetings with instructional staff. Program directors who complete elements one through three are eligible to remain in the PBAS and therefore eligible for funding from the county (per child, per day reimbursement). Organizational instability is defined as manager turnover, turnover of over 50 percent of staff, change of program location or other catastrophic service interruption.

[19] The per-child reimbursement in this case is directly tied to attendance of qualified youth using a per child per day rate from county. However, for a program to qualify for the reimbursement scheme, they must participate in the QIS so the connection between QIS participation and per child reimbursement is mediated by attendance.

Following a *baseline* year PBAS cycle for quality management practices, programs move to the *intermediate* level in all subsequent years (no longer required to complete a program self-assessment, which is only required during the first program cycle). With specific quality management practices in place,[20] when programs attain a program rating (PBC-PQA Form A) for quality instructional practices greater than 4.1 for two successive years, they move to the highest level, *maintenance*. Beginning the 2013 cycle with 121 programs, 16 were at the baseline level, 82 were at the intermediate level and 23 were at the maintenance level of the PBAS.

Discussion of the Palm Beach County PBAS With Reference to the PBAS Design Standard (Figure 1)

With this summary in hand, we can discuss the Palm Beach County PBAS in terms of the design, components and linking supports defined in Figure 1.

Quality is a broad public goal. One of the most critical characteristics of the Palm Beach County PBAS was the focus on the measurement of core setting processes during the formative years of policy roll-out. Several implications are critical to understanding this decision. First, because the core service outputs are processes – quality management and quality instruction – that aim to deliver high quality child experiences at the point of service, service quality has itself been a publicly stated goal for the PBAS. While child outcomes are clearly rising to the top of the list of publicly stated goals in recent years, stakeholders continue to describe access to high quality experience as a key output of the system, and communications to parents are framed in terms of service quality.

A second reason for the designation of an "output" as a broad public goal is simply because a well-specified logic model (Figure 1, Service Production Map) was not available in Palm Beach County or elsewhere in the field. The Palm Beach County logic model contains considerable detail on the "left hand" side, where production inputs and outputs are described.[21] However, the right-hand side of the model, including individual level outputs and broader outcomes (e.g., 21st century skill development; literacy skills) is less well-developed, reflecting a lack of clarity in both the science and the affiliated professions.

The Palm Beach County PBAS emphasizes performance improvement supports and methods as part of its PBAS design (Figure 1, Performance Improvement Map). The Palm Beach County PBAS clearly articulates organization-level quality management practices and quality instructional practices, and

[20] For a program to move to the maintenance level the *Director Action Plan* requires a program director to provide artifacts and check-in's with Quality Advisors as evidence of the following quality management practices: (1) complete satisfactory improvement plan with SMART goals; (2) complete and document satisfactory observation-feedback sessions with staff; (3) quarterly staff meetings include a focus on improvement plan goals; (4) complete satisfactory progress checks with focused on the improvement plan.

[21] An extension of the theory of change to include individual skill development can be found in the Feasibility Study for Impact Evaluation and Intervention Design Improvements (Smith, Akiva, Gersh, & Sutter, 2012).

makes evidence-based training and technical assistance available to assure implementation of these core service outputs. Furthermore, the process of measuring the implementation of quality management practices and quality instructional practices is located in an external "third party" organization (Family Central), so that disputes over performance measurement are distinct from access to training and technical assistance or from links to incentives.

The PBAS design differentiates by management and delivery roles. The Palm Beach County logic model describes service outputs at multiple levels of setting: organizational stability and quality management practices at the organizational level, and quality instructional practices at the offering session level. This multi-level definition of service outputs supports differentiation by roles in the PBAS design. Program managers working to implement quality management practices and program teachers implementing instructional practices are differentiated across the components and linking supports in Figure 1, specifically in terms of measures, ratings and performance levels, incentives, and training and technical assistance for improvement.

The Palm Beach County PBAS has evolved from lower to higher stakes with an early emphasis on technical supports and quality management practices and with later emphasis on financial rewards and identification of higher performers. The PBAS design has been evolved skillfully over time by leadership from an overall, lower stakes design to one that includes higher-stakes elements. Initially, the PBAS emphasized implementation of quality management practices (YPQI) which is a lower stakes accountability because implementation can be successfully completed by almost all program managers and because those program managers report that implementation of quality management practices is a good use of their time (Sugar, Pearson, Devaney, & Smith, 2009). In the early years, the only incentives attached to program ratings for quality management practices or quality instruction practices were access to improvement supports. Presumably, the lower performing programs received "rewards" in the form of training, technical assistance, and coaching to improve performance.

As quality management practices were implemented at higher fidelity over successive PBAS cycles, program ratings describing the quality of instruction began to improve, as suggested by the logic model for service production in Figure 2. With a program management workforce equipped to implement quality management practices (e.g., Palm Beach County's YPQI), completion of these practices were tied to per-child funding. As program ratings for quality of instructional practices moved over a designated performance level (program quality rating of 4.1 or greater), the program rating for instructional quality became an accountability in PBAS with an associated performance level and specific incentive (public recognition). In this case, program staff were not threatened by the new accountability because the performance level was attainable and because the incentive emphasized good performance only.

Part III. Using Palm Beach County Data to Examine Reliability of Program Ratings and Validity of PBAS of Performance Levels

Background

Because youth experience is determined in important ways by the qualities of the settings in which they spend their time, one of the central concerns in the expanded learning field is to determine "How much quality is enough?" This question raises several implications – chiefly, that we can identify an optimal experience for youth, then use funding and accountability resources to distribute that optimal level of youth experience for maximum social return. More specifically, if we can determine a threshold for quality where either more quality does not add value in terms of youth development, or a low threshold under which negative consequences for youth might occur, we can target funding and accountability resources toward getting all lower scoring programs to the minimum optimal level. We also might redistribute some resources away from the programs above the high threshold.

While the issue of thresholds for quality is primarily a concern for developmental science, because we do not yet know what thresholds for youth experience might be, there are clearly threshold-like issues that influence the design of PBAS policies in the expanded learning field. Decisions about the composition of quality ratings and the setting of performance levels in PBAS are based on hunches about quality thresholds that are yet to be validated by developmental science. The challenge: Developmental science will require time and funding for rigorous studies to produce evidence about thresholds, while performance levels are part of an emerging performance management paradigm that requires decision-making in the present.

Insights on Quality Thresholds From Research in Early Childhood Education

Review of findings from early childhood research on developmental thresholds. A review of research on quality thresholds in the early childhood field is available (Zaslow et al., 2010), as is an exemplary study with methods that could be directly applied in the expanded learning field (Burchinal, Vandergrift, Pianta, & Mashburn, 2010). The Zaslow, et al., review provides insights from studies that provide rigorous tests of hypotheses about thresholds using large samples drawn from typical preschool classrooms with repeated measures for service outputs that include both process quality[22] and change in individual skills. Key points from the review and a few other relevant studies include:

- Past research suggests that measures of structural features which produce naturally occurring thresholds (e.g., teacher has bachelor's degree or not; program is accredited or not) are

[22] In this sub-section we use the term *process quality* which is the early childhood equivalent for the term *instructional quality* used throughout the rest of the paper.

inconsistently related to child outcomes. Following a logic model for service production, service inputs (i.e., structural features) like staff education, training and experience are not very convincingly related to service quality in the early childhood literature (Early et al., 2007; Mashburn et al., 2008). In the evaluation of the Qualistar PBAS for early childhood settings in Colorado, there were also no relationships between any measures for teacher education, training[23] or experience, and only teacher experience was related in any way to process quality (Early et al., 2007; Zellman et al., 2008). It is also clear that these inputs are probably not measured well, reducing our confidence in results. Researchers suggest that structural features are important as predictors of process quality, and, in turn, that process quality is the key predictor of child-level change (Mashburn & Pianta, 2010). Compelling evidence suggests that, while rarely measured, organizational-level processes like quality management practices are probably more proximal predictors of process quality (Smith, Akiva, Sugar, et al., 2012).

- Process quality (e.g., adult-child interaction, instruction) is consistently associated with child-level change, but the thresholds for change are not clear, i.e., there is mixed evidence of non-linear relationships that imply the presence of a threshold. There is very little evidence to suggest a high threshold for observed quality, beyond which the effect on children is reduced. The best evidence available suggests a moderate-to-high threshold for quality below which little effect on children is achieved, but above which additional outputs of quality produces additional outputs of individual skill growth (Burchinal et al., 2010).[24] Developmental theory, in our interpretation, argues against the concept of reaching an upper limit to the effectiveness of a setting where adults are working with groups of children (Fischer & Bidell, 2006).
- Making the situation much more complicated, there is some evidence that thresholds differ by student need (i.e., students who have fewer supports for self-regulation at home may benefit more from the experience of high quality than peers with stronger home supports) and by content area (i.e., threshold scores may differ for social and emotional experiences, compared to meta cognitive experiences).

[23] There is little evidence that stand alone training experiences have much of an impact on anyone. This is probably even truer of the ubiquitous "attendance at conference workshops."

[24] The issue of avoiding harm is also relevant here since the literature from early childhood presents a decidedly mixed record of effects, especially for social and emotional learning. While the very high quality and very tightly controlled interventions – Perry Preschool, Abecedarian, Chicago Child and Parent Centers – have produced a consistent record of positive and long-term effects, more run-of-the mill early childhood experiences have produced a more mixed record with negative effects on social and emotional learning measures in some studies (Baker, Gruber, & Milligan, 2008; Lowenstein, 2011; McCartney et al., 2010), which are likely attributable to low quality child experience. While the after school research literature has no comparable set of studies with measures of both program quality and pro-social behavior, there is some evidence suggesting that after school programs can have negative effects on school behaviors for some students (James-Burdumy et al., 2005).

Performance Levels in Early Childhood PBAS. The convergence of developmental science, applied measurement, and performance levels, and public policy is demonstrated in this excerpt from the leading researchers on this issue in the early childhood field:

> A primary goal has been to identify levels in the association between quality and child outcomes at which the linear association begins to asymptote off, above or below which there is little evidence of increases in learning associated with increases in quality. A threshold that indicated that the quality-outcome association levels off asymptotically above a given level of quality would suggest that policies should focus on improving quality up to that threshold level, but improving quality above that point may not be necessary for improving child outcomes. Policy to address that goal would invest in lower or average quality classrooms while leaving classrooms with quality above the threshold alone. In contrast, it is possible a threshold could define the minimum level at which a positive association between quality and outcomes is observed. In this scenario, there may be no detected relation between quality and outcome gains until quality reached a certain point on the scale; in other words, learning did not take place until classrooms demonstrated a minimal level and after that minimum, gains in learning increased as quality increased. This form of threshold effect would suggest that it is especially important to ensure that children experience at least the minimum level of quality child care in order for those experiences to be related to improved child outcomes. It would point perhaps to not allowing vouchers to pay for care that was below the threshold, while also incentivizing teachers above the threshold to continue to improve. (Burchinal et al., 2010, p. 167).

To date, performance levels for early childhood quality measures have been based on expert guidance from the instrument developers rather than empirical validation. High and low performance levels on the Early Childhood Environmental Rating Scale (Howes, Phillips, & Whitebook, 1992) and the Classroom Assessment Scoring Systems (La Paro, Pianta, & Stuhlman, 2004) are both defined as: 1-2 is low, 3-5 defines a mid-range, and 6-7 is considered high. However, in most studies that use these measures, performance levels are selected based on sample variation, so the developmental research is often not designed to validate specific performance levels that might be selected for PBAS.

Study Questions

The empirical findings discussed in this section address questions about the reliability of program ratings and the validity of performance levels designated in the PBAS. In this section, we focus only on program ratings for quality of instructional practices composed from the PBC-PQA Form A scores. The presentation of analyses and results is organized in three steps. First, we examine the precision of program ratings in terms of reliability and sensitivity. Next, we conduct several preliminary validity analyses related to PBAS performance levels. Finally, we execute a Rasch measurement model to identify

"naturally occurring" performance groups while controlling for activity type. Specific study questions include:

- Does the composition of scale and offering session scores on the PBC-PQA Form A produce program ratings that are sufficiently reliable to support several PBAS uses, including classification of programs by level of performance?
- Are there ceiling effects for the program rating scale that limit the ability of the PBAS to classify programs at high levels of quality?
- Does the type of offering session content influence the program rating in a way that could influence assignment to a performance level in the PBAS?
- What is the effect of selecting different performance levels on (a) the distribution of programs within each level and (b) on the magnitude of rating differences across performance levels?
- Can we identify individual performance subgroups that provide empirical justification for selection of program performance levels?

Program Characteristics and Instructional Practice Measures

Program Characteristics

In a 2012 survey of 83 program leaders, Palm Beach County PBAS programs were described as a mix of community-based (37.3 percent) and school-based (62.6 percent) providers that offer after chool services during the school year, frequently offer summer sessions (68 percent) and serve between eight and 200 children/youth per day. All programs offer services to elementary-aged children, while some also provide services for middle school (21 percent) and high school (11 percent) youth. Nearly all (94 percent) programs are open to the general child/youth population, while some also target low-income students (58 percent), limited English speakers (27 percent), students in foster care/child welfare situations (23 percent), students with physical or learning disabilities (21 percent), students in migrant families (13 percent), and students whose families have immigrated from outside the US (12 percent) (Smith, Akiva, Gersh, et al., 2012).

Instructional Practice Measures

Measures. The Palm Beach County Program Quality Assessment[25] (PBC-PQA) is the measure for quality of instructional practices in the Palm Beach County PBAS. For these analyses, we composed

[25] The PBC-PQA is based on the Weikart Center's Youth Program Quality Assessment (Youth PQA) Form A, with only a few minor differences. However, because of these differences, scores for the PBC-PQA are expected to be marginally higher than scores on standard Youth PQA because, for example, the PBC-PQA does not include a few of the lower scoring items in the Engagement scale. The methodology for completing the PBC-PQA is the same as the Youth PQA (Smith & Hohmann, 2005).

program ratings according to the following rules: We used 50 of the PBC-PQA items that nest within 15 scales. An instructional total score for each observed offering session[26] was created as the unweighted average for the 15 scales. The program rating was then created as an average of the instructional total score for each of three program offering sessions. Again, each of the three instructional total scores was produced during sessions from different offerings with different staff members so the program rating includes an observed sample of instructional practices from at least three different staff. Appendix E and Figure E-1 describe the composition models for the instructional total score for an offering session and for the program rating.

Data collection methods. Each year programs participating in the PBAS are visited by endorsed external raters from a third-party assessment organization contracted by Palm Beach County Children's Service Council.[27] External raters are endorsed by passing video tests for item-level reliability that requires perfect agreement with "gold standard" scores on 80 percent of PQA items. Raters undergo annual paired-rater tests and quarterly checks to remain endorsed. At the beginning of the year, program directors are notified about the month during which the observations will occur and then receive 24- to 48-hour notice before the external rating visit begins. Three offerings are selected randomly on the day when the data collector arrives at the site.

Data file. The program ratings data file includes all program sites participating in the Palm Beach County PBAS from 2008 through 2012. Table 3 provides the number of program ratings and the number of offering session ratings used to create the overall program ratings for each year. The largest increase in participation was in 2009, when the number of program sites increased from 64 to 90, and in 2011, when the number of sites increased from 93 to 114. Forty-two afterschool program sites have quality ratings for all five years.

Table 3. Number of Program Ratings and Total Offerings Observed

	2008	2009	2010	2011	2012
Number of Program Ratings	64	90	93	114	115
Number of Offering Session Ratings	192	270	279	342	345

Rasch scores. Most of the analyses presented are conducted using unadjusted measures for instructional practices. In several analyses, however, we calibrate PBC-PQA scale scores using Rasch

[26] An offering session is defined as a single instance of the same group of children and the same adult who meet for a named learning purpose (e.g., hip hop dance; math club) over multiple sessions.

[27] Importantly, the Family Central organization is also the third-party assessor for the early childhood system and one of the few examples of assessment capacity in the same organization selling services to independent federal and state funding streams. Presumably, this is a capacity that could be built in many cities and states if the markets for assessment in early childhood, expanded learning, and the school day were identified.

methods.[28] Rasch modeling techniques yield both a program rating and an estimate of the difficulty of each PBC-PQA item. Rasch techniques transform program ratings and item difficulty estimates using a log function, allowing program ratings and item difficulty estimates to be arrayed and compared along the same logit scale.

Findings for Reliability and Sensitivity of Program Ratings

Before considering PBAS performance levels and the validity of these levels and their use, it is important to have some understanding about the precision of the program ratings in terms of reliability and sensitivity. These facets of precision are of particular interest because PBAS are designed to differentiate between programs at given points in time and to track performance change over time.

Reliability of Program Ratings

Reliability can be understood as the degree of consistency or agreement between a set of indicators or ratings. In this section, we examine (a) the consistency of 15 PBC-PQA scale scores as components of an instructional total score for an offering session, (b) the consistency of three offering session instructional total scores as components of a program rating, and (c) the consistency of annual program ratings over multiple years.

Internal Consistency of 13 Scales Composed as the Instruction Total Score for An Offering Session. One key attribute of the reliability of the instructional total score is the consistency with which the 15 component scale scores indicate high and low levels of instructional quality. In this sense, each scale is considered a nonindependent observation of instructional quality during an offering session. We calculated the internal consistency for the instructional total score for each of the five years for which we have data, and Table 4 presents these results. The average internal consistency coefficient for the instructional total score is $\alpha = .80$[29]. In general, alpha coefficients of .7 or greater are considered acceptable (Nunnally, 1978).

Table 4. Internal Consistency for the Instructional Total Score

	2008	2009	2010	2011	2012
Cronbach's Alpha	.78 (N=192)	.82 (N=268)	.78 (N=279)	.80 (N=339)	.80 (N=339)

[28] For an extensive treatment of reliability and validity issues for the School-Age Youth Program Quality Assessment using Rasch modeling techniques, see Naftzger (2012).

[29] It is worth noting that we consider the item level of measurement for the PQAs to be formative rather than reflective (Diamantopoulos, 2008). The reliability of item level measurement on the PQA is best assessed as inter-rater and test-retest reliability at the item level and has been addressed elsewhere (Smith, Akiva, Sugar, et al., 2012). For this study, item-level reliability was addressed through rater training as described in the measures section of this report.

Reliability of Program Ratings. In the Palm Beach County PBAS, program ratings for a given year are composed as an average of three instructional total scores for sessions from three different offerings. In this case, reliability for a program rating is defined as a high degree of agreement between the three instructional total scores used to compose that program rating.[30] To the extent that a program rating consists of instructional total scores that are very different, the rating will inadequately serve its primary PBAS purpose: to represent the quality of instructional practices available at a program in a straightforward way.

We used two types of intraclass correlations (ICC[1]) and ICC[2]) to describe reliability as agreement for the program ratings.[31] Using the statistical software package Hierarchical Linear Modeling (HLM), an unconditional HLM was used to estimate variances for the 2010 year of Palm Beach County PBAS ratings. An estimate of the rating variance between programs (tau) is divided by an estimate of the total variance for the rating sample, which includes both the betwee- program estimate and an estimate of rating variance within programs (sigma squared). The intraclass correlation coefficient (ICC[1]) for the Palm Beach County year three data is 0.07/(0.07 + 0.16) = 0.44.[32] This coefficient can be interpreted in two ways following Bliese (2000). First, roughly 44 percent of the variance in offering sessions is attributable to differences in programs. This means that a substantial amount of the variation in the quality of instruction during youth offerings is attributable to characteristics of the program itself, providing some justification for the construction of a program rating. As a comparison, in the evaluation of the Qualistar early childhood PBAS in Colorado (Zellman et al., 2008), similar ICCs were estimated for the Early Childhood Environment Rating Scale, indicating that nearly 70 percent of rating variance was attributable to programs. Second, this coefficient can be understood as the reliability of a single rating of an offering session as a representation of overall program quality. In this sense, the coefficient would be considered low, and for this reason the PBAS requires multiple ratings for each program.

Using the ICC[1] coefficient, it is also possible to account for the number of ratings per program and estimate the reliability of the overall program rating consisting of a given number of observations (ICC[2]). Again following Bliese (2000), we apply the ICC[1] in the following formula, Reliability = k(ICC) /1+(k-1)ICC, where k= the sample size in each program rating. In this case, k=3 instructional total

[30] This type of reliability is grounded in a literature that describes the reliability of *group means* and typically examines ratios of score variances for different components of a quantitative model (Brennan, 1995).
[31] Much more fully specified measurement models are possible. Our approach, which largely ignores error associated with raters or items, is described as approach A in comparison to other more fully specified models in Schweig (2013).
32 A confidence interval for the ICC was estimated for the Palm Beach County year three data of 0.42-0.46 (Raykov, 2013).

scores for each program. The resulting reliability coefficient (ICC[2]) for a program rating is .70[33]. ICCs greater than .70 are generally considered an indication of acceptable levels of agreement, although there is much debate about the use of arbitrary cutoffs for this index (Harvey & Hollander, 2004; James, Demaree, & Wolf, 1993). For example, findings from program quality studies in the medical field suggest that ICCs as low as .40 are interpretable (Haut et al., 2002).

Because group reliability estimates are difficult to interpret, we also estimated confidence intervals for each program rating produced in the 2010 program cycle using one-way analysis of variance. The average confidence internal for all 93 program ratings in 2010 was 1.48 scale points (i.e., plus or minus .74 scale points on average), indicating the relative imprecision of the PBC-PQA. Because the Prime Time PBAS is focused on high scores, we also estimated the confidence interval for the top quartile of program ratings in 2010 which was .89 scale points.

Stability of Program Ratings Over Time. The final aspect of consistency we considered is the stability of ratings over time. Stability is a tricky issue in this case because we expected scores to rise over time due to participation in the PBAS. In effect, we hypothesized that ratings would be moderately correlated in contiguous years and less highly correlated in subsequent years because scores should rise over time. Table 5 provides bivariate correlation coefficients for program ratings over five years. Means and sample sizes are presented parenthetically in the column headers. In general, correlations are moderate to small in magnitude, ranging from .20 to .59. The average correlation across all years in Table 5 is .36 while for successive years is the average correlation is .44. Because reliability of the program rating is known (.70), it is possible to estimate these correlations with adjustments for unreliability or error in the overall rating (Wang, 2010). With this disattenuation (divide the correlation coefficient by the pooled reliability of the two measures or r=.44/.70), the average correlation coefficient across successive years is .63. As a point of comparison, unadjusted correlation coefficients for a similar observation-based rating for preschool classrooms, the Classroom Assessment Scoring System (La Paro et al., 2004), were .59 for the Emotional Support scale and .32 for the Instructional Support scale (Burchinal et al., 2010, p. 169).

[33] Calculated for each of the other years in the data set the comparable ICCs are: 2008=.63, 2009=.56, 2011=.63, 2012=.70. Calculated for the entire sample of 1,428 offering session ratings collected over five years and nested within 143 programs, 40 percent of the variance in offering session ratings is explained by program and the sample adjusted reliability coefficient is .77.

Table 5. Program Rating Correlation Matrix Over Five Years

	2008 (N=144)	2009 (N=127)	2010 (N=93)	2011 (N=70)
2009	.50**			
2010	.31**	.30*		
2011	.31**	.40	.59**	
2012	.20	.25	.35**	.36**

**. P< 0.01 level (2-tailed).*. p< 0.05 level (2-tailed).

Offering Type as a Covariate in Rating Measurement Models. The relationship between offering type and offering session scores on the Youth Program Quality Assessment (Smith & Hohmann, 2005) has been replicated in several samples (Akiva, Smith, Sugar, & Brummet, 2011; Naftzger et al., 2012; Smith, Peck, Denault, Blazevski, & Akiva, 2010). One major challenge in collecting data on the quality of instructional process is that offering type is not well-defined, resulting in coding schemes that mix content (e.g., life skills, STEM) and pedagogical approaches (e.g., enrichment, tutoring). Offering type could introduce systematic bias into program ratings if, for example, one program rating consisted of scores for three homework help offerings and another program rating consisted of scores for three academic enrichment offerings. For this reason, offering type can be understood as a threat to rating validity and a potential source of bias.

We used the Palm Beach County offering-level data file to examine the effect that offering type might have on program ratings in the Palm Beach PBAS. Table 6 provides the mean and standard deviation for six offering types. Independent sample T-tests were conducted for each offering type in comparison with the academic enrichment activities which are a normative type of offering across different coding schemes, and typically the highest scoring type of activity. Homework, tutoring, sports and story time offerings scored lower than academic enrichment, while offering scores for arts and crafts were similar to academic enrichment.

Table 6. Mean Differences in Offering Ratings (Instructional Total Score) by Offering Type, All Years, N=1146 (282 missing)

	Homework (n=87)	Tutoring (n=5)	Academic Enrichment (n=378)	Arts and Crafts (n=356)	Sports and Games (n=307)	Story Time (n=13)
Mean for Instructional Total Score (SD)	3.66 (.47)	3.37 (.47)	3.91 (.47)	3.87 (.51)	3.67 (.51)	3.67 (.26)
Difference from academic enrichment score is statistically significant $p \leq .1$	Yes	Yes	NA	No	Yes	Yes

Sensitivity

A reliable program rating can be located on the quality scale at a point relatively near actual performance, so that programs can be consistently differentiated on the basis of their rating. Sensitivity refers to the usefulness of the scale for reflecting differences in ratings, especially change over time. Specifically, we want to know if the quality scale has room for most programs to grow, since the purpose of participation in a PBAS is premised on improvement of service quality. Ceiling effects limit the usefulness of a scale for the PBAS because evidence of improvement cannot be linked to incentives if change cannot be captured and represented on the scale.

In general, change scores for Youth PQA total scores have been under 1-scale-point per cycle of YPQI implementation. For example, in the YPQI study assignment to the group implementing YPQI in their programs produced an average effect of approximately one-quarter of a scale point for all programs. High implementation of YPQI was associated with three-quarters of a point improvement on the Youth PQA scale (Smith, Akiva, Sugar, et al., 2012). In a review of year-to-year change in program scores from several cities with YPQI implementations, Youth PQA total score ratings increase, on average, by four-tenths of a point on the rating scale during the first year and one-10^{th}-of a point in subsequent years (David P. Weikart Center for Youth Program Quality, 2012, 2013a, 2013b)

In order to better understand the sensitivity of the PBC-PQA rating scale, we used Rasch measurement methods to examine the distribution of 1,429 program ratings in the Palm Beach County PBAS data set. Rasch scaling allows presentation of program ratings and item difficulty estimates on the same scale and the resulting "variable map" is provided in the first section of Appendix F. In general, these analyses suggest that the program rating scale does not have serious floor or ceiling effects. However, it is clear that the PBC-PQA has many "easy" items on which nearly all programs score high. While important as standards for performance, these items do little to differentiate between programs because almost all programs attain the highest score.

Findings for Exploratory Analyses for PBAS Performance Levels

Thinking back to elements of Figure 1, the previoussection described PBAS measures and their composition. This section will examine several aspects of PBAS performance levels. Again using data from the Palm Beach County PBAS, we first examine the program ratings for all programs, on average, in each of the five years of PBAS data, then track selected individual programs' ratings through time. Our purpose is to gain some understanding of how well the PBAS data reflects the actual dynamic purposes of the PBAS (for scores to improve over time) and to gain some perspective on how much scores might increase. This kind of information is critical for calibrating performance levels to fit actual patterns of performance improvement. These analyses can also be considered an examination of the validity of the PBAS in the sense that we are asking if PBAS improvement methods produce intended outputs, and if PBAS measures and performance levels can represent improvements that occur.

Next we examine the effects of using different sets of performance levels on the classification of programs. Different assumptions about high and low performance produce different results in the classification of programs and importantly, in distribution of incentives. Calibration of PBAS performance levels to achieve a good distribution of incentives requires an understanding of classification effects produced by assumptions about high and low.

The second section of Appendix F uses Rasch modeling techniques to explore naturally occurring performance subgroups for individuals that might inform decisions about performance levels in organizations.

PBAS Program Performance Levels and Trajectories

The Palm Beach County PBAS is designed to improve the performance of programs delivering expanded learning services. One critical output in the production chain for these services is the quality of instruction. The PBAS has aligned a measure to this part of the process, the PBC-PQA. The total score for three PBC-PQA ratings are averaged together to produce an annual program rating. The first three columns of Table 7 provide mean program ratings for all PBAS programs in each of the five years in the data file. These scores present a substantial increase from 2008 through 2012, as well as increases in all but one of the four possible consecutive year combinations. However, there are several challenges with this view: (1) new programs enter across time, meaning that each year's mean rating does not represent the same group of programs; (2) these are aggregate ratings that may mask actual patterns of change within a specific program across years; (3) the increments of change year-to-year are small.

The last column in Table 7 addresses the first issue by selecting the 42 programs that have participated in the PBAS for all five years. Here, the scores increase in each successive year, although in very small increments in some years.

Table 7. Mean Ratings for All Programs and Five-Year Programs

	Rating Mean (SD), All Programs	Rating Range, All Programs	Rating Mean, Five-year Programs, N=42
2008	3..49 (..40) N=64	2.38 - 4.62	3.47
2009	3.79 (.36) N=90	2.49 - 4.65	3.80
2010	3.90 (.35) N=93	2.86 - 4.65	3.92
2011	3.81 (.37) N=114	2.98 - 4.78	3.89
2012	4.01 (.37) N=115	2.97 - 4.83	4.06

While scores for the five-year participants increase in small increments, the ratings are at the high end of the PBC-PQA rating scale. So it is possible that these programs are "high enough" and further increase is either not possible or not their intention. For example, the Palm Beach County PBAS has designated a rating of 4.1 for two successive years as a key performance level (see section 2). Two further questions relevant to the validity of the Palm Beach County PBAS are: Do programs that attain the high performance level sustain that level of performance? Do programs that start low increase over successive years?

To address these questions, we divided the program data file for the 2009 year into three groups of programs: The high quality group consisted of programs with ratings one standard deviation above the mean rating or higher (15 percent); the mid-quality group consisted of programs with ratings within one standard deviation above or below the mean (70 percent); the low-quality group consisted of programs with ratings one standard deviation below the mean rating or lower. We then gathered four years of data, 2009-2012, for the 10 highest scoring programs in the high-quality group (ratings 1 SD or more above the mean 2009 rating) and for the 10 highest scoring programs in the low-quality group. In order to sustain a group of 10 within the performance level cut point, some groups with fewer than five data points were admitted to the sample.

Figure 3 presents rating trajectories over four years for the top scoring programs in the low-quality group. From Figure 3 several points are evident. First, the mean rating for these programs goes up over time as denoted by the heavy dashed line (exceeding the magnitude of the sample standard deviation), and the general pattern is an upward trend in nearly all trajectories (not a few dominant

outliers with sharp upward slopes). Second, there is some year-to-year volatility in the ratings, with 40 percent of the programs (23, 42, 70, 146) experiencing a decline in ratings in one successive year period but only one program (204) experiencing a decline in ratings for two successive years. Third, the two programs that exceed the Palm Beach County PBAS performance level of 4.1 (108, 144) in 2010 maintained or exceeded that level for the remaining years.

Figure 4 presents rating trajectories over four years for the top scoring programs in the high-quality group. From Figure 4, several points are evident. First, the mean rating for these programs stays almost the same over time, as denoted by the heavy dashed line. Second, there is some year-to-year volatility in the ratings, with nine of 10 of the programs experiencing a rating decline in at least one year, but only one program experiencing a decline in two successive years (19). Third, all 10 programs began above the Palm Beach County PBAS performance level of 4.1 and only two of the 10 programs dipped below that level (14, 125), although both of these programs were above the 4.1 performance level in all successive years.

Figure 3. Four-Year Trajectories for Ten Programs Scoring Low in 2009

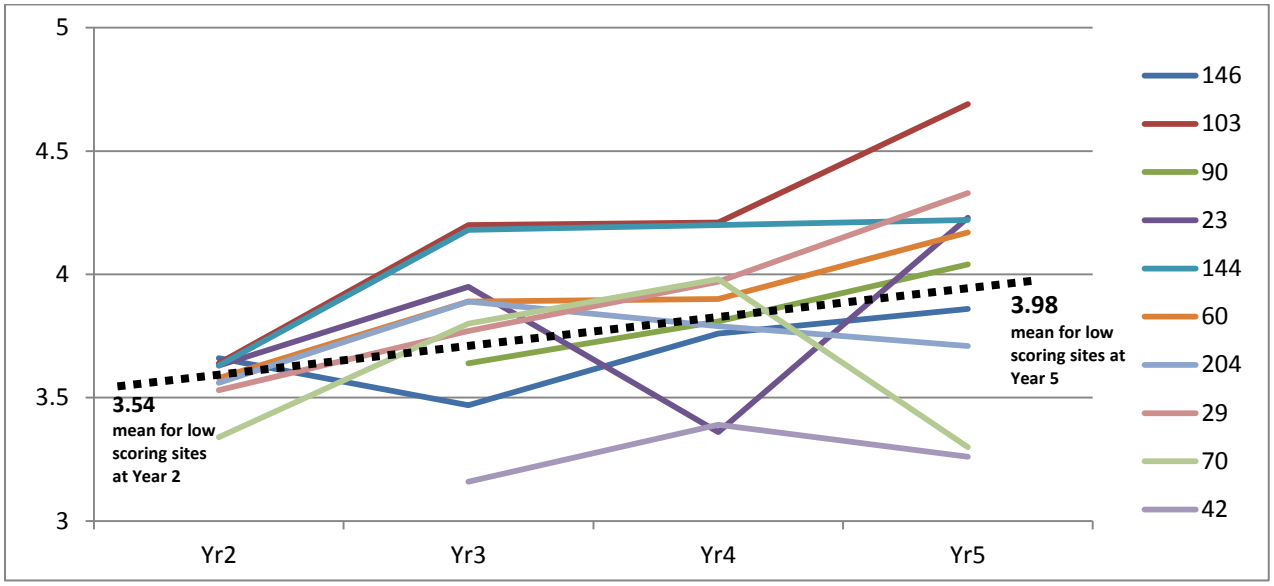
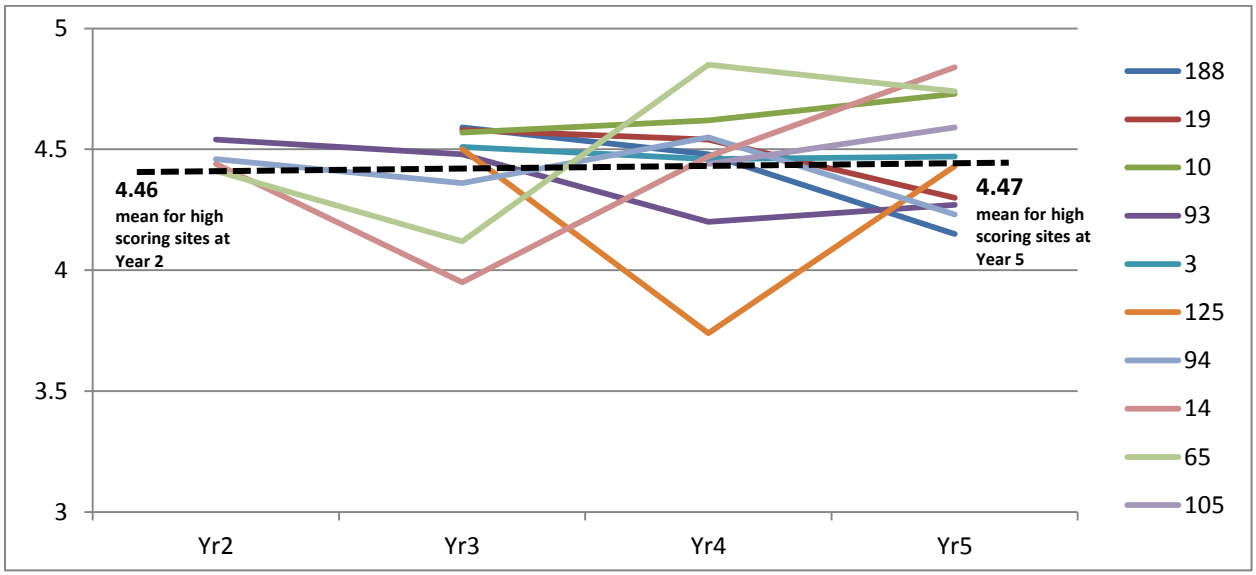


Figure 4. Four-Year Trajectories for Ten Programs Scoring High in 2009



Classification of Programs Using Different Performance Levels

Program Ratings Simulation. Different assumptions about high and low performance produce different results in the classification of programs and, ultimately, in the distribution of incentives (Figure 1, c) and supports (Figure 1, f) associated with a particular level of performance. To better understand how different decisions about performance levels would affect the classification of programs, we used the 2010 data from the Palm Beach County PBAS for purposes of simulation. Table 8 provides descriptive detail for four different methods of classifying programs into groups or bands of performance: none, median, quartiles, standard deviation, Palm Beach County PBAS. Each of these schemes produces different sets of performance bands (e.g., 2nd quartile group). For each performance band, the table provides the mean rating and standard deviation, the proportion of total 2010 programs in that performance band, and a statistical test for mean difference between the means for the lowest and highest bands.

Table 8. Descriptive Findings for Different PBAS Performance Level Configurations Using Program Ratings for 2010, N=93

Method	Definition of Level	Mean Rating (SD) for Level	Proportion of Sites Assigned to Level	Statistical test for different between highest and lowest category
All sites	None	4.02 (.32)	100%	NA
Median	Below Median Above Median	3.76 (.20) 4.27 (.19)	50.5% 49.5%	t(91) = -12.69, p=.000
Quartiles	Q1 (low) Q2 Q3 Q4 (high)	3.62 (.19) 3.91 (.05) 4.12 (.07) 4.43 (.13)	26.4% 24.2% 25.3% 24.2%	t(44) = -16.86, p=.000
Stand Dev	1 SD below (low) *0.00 ≤ n ≤ 3.70* Within 1 above or below *3.71 ≤ n ≤ 4.33* 1 SD above (high) *4.34 ≤ n ≤ 5.00*	3.50 (.18) 4.01 (.17) 4.50 (.09)	14.0% 69.9% 16.1%	t(26) = -18.53, p=.000
PB Method	Baseline Intermediate Maintenance	NA NA 4.1	13% 68% 19%	NA

Part IV. Discussion of Findings and Recommendations

> "[A] framework provides a foundation for descriptive and prescriptive inquiry by establishing a set of assumptions, scope, and general classifications and relations among key concepts" (Weible, Sabatier & Kelley, 2009).

The field of expanded learning has grown quickly in size and sophistication in recent years. The progress to date has been scaffolded in part by performance-based accountability systems (PBAS) that are at the leading edge of policy implementation and intervention science. The specification and description of these systems has also made substantial progress with field-level reviews of system designs (Yohalem et al., 2012; Yohalem, Ravindath, et al., 2010), quality intermediary organizations (Collaborative for Building After-School Systems, 2007), and core output measures (Wilson-Ahlstrom et al., 2011; Yohalem et al., 2009). The nomenclature and framework for performance-based accountability systems developed by Camm and Stecher (2010) advances our thinking and ability to describe expanded learning systems that join disparate actors in the pursuit of shared community goals for children and youth.

Many communities around the United States are engaged in collective impact projects that center on the assembly of dashboards to display various types of data. There is little evidence, and indeed little logic, suggesting that this kind of information has a direct effect on performance change (Aguinis, Joo, & Gotfredson, 2012; Weitzman, Silver, & Brazill, 2003). This is true because for individuals, insight is weakly correlated with action – or in more specific terms, data alone do not typically support a knowledge management sequence where the data are converted to meaningful information that is shared by a professional community and that can be acted upon to develop expertise about practice (Halverson, Grigg, Prichett, & Thomas, 2007; Mason, 2003). The expanded learning field has built the PBAS capacity necessary to manage individuals and organizations toward successful attainment of performance goals, fulfilling the adult learning purposes that should be at the core of accountability policies.

In the next two subsections we present a discussion and findings. We first focus on the usefulness of the PBAS framework for describing an expanded learning system in Palm Beach County. Next, we review technical issues regarding program ratings and performance levels related to a critical service output, the quality of instructional practices. In the final subsection we present recommendations for Weikart Center clients and others that (a) use the Youth Program Quality Intervention to define a set quality management practices and (b) use the Youth or School-Age Program Quality Assessment to define a set of quality instructional practices.

Discussion of Findings

Findings from Application of the PBAS Framework

In section II and Figure 1 we adapted the PBAS framework and nomenclature by highlighting and integrating existing performance management capacities in the expanded learning field. Specifically, we add four linking supports to the primary PBAS components to define a system design standard that includes: (a) goals, (b) a map for service production, (c) a set of measures, (d) a map for composing performance ratings and setting performance levels, (e) a set of incentives, (f) a performance improvement method, and (g) effective performance data that supports the improvement method. We used one of the nation's most mature expanded learning systems – in Palm Beach County, Fla. – as an exemplar of best practice, to see how well the PBAS framework could describe the Palm Beach County system.

Our findings from this process of fitting the PBAS framework to an expanded learning exemplar include the following:

- Quality improvement systems in the expanded learning field are performance based accountability systems, and the PBAS framework can be extended for applications in the expanded learning field with inclusion of several linking supports.
- As a test of the robustness of the PBAS framework, the Palm Beach County PBAS can be more fully described, and with more technical specification, using the PBAS design standard described in Figure 1.
- Several important attributes of expanded learning PBAS were clarified in the process:
    - Measurement of service outputs is the central task of a PBAS. Important service outputs include both the quality of processes at multiple levels (management practices, instructional practices) and individual skill trajectories (21st century skills, school success).
    - An effective PBAS differentiates by organizational roles across components and linking supports in order to heighten the effect on individual behaviors of managers and teachers.
    - While many measures can be aligned with the active ingredients identified by the service production map, attention to the composition of the measures into program-level ratings of known reliability is important.
    - Performance levels define good performance and are required for clear alignment with incentives. Performance at any specific program can be defined by multiple program ratings (e.g., program stability, quality management practices, quality instructional practices, youth engagement) and corresponding performance levels for each rating scale.

- Higher-stakes incentives include:
  - Amounts of funding that could substantially affect program operation.
  - Forms of external recognition that could negatively affect program enrollment (sharing low ratings publicly).
  - Forms of internal recognition that could negatively affect the reputation of individuals (sharing low ratings within system or program staff).
- Lower stakes incentives include:
  - Access to performance data.
  - Access to training and technical assistance for improvement methods.
  - Amounts of funding that cannot substantially affect program operation.
  - Forms of external recognition that identify high performance.
- Lower stakes PBAS designs that emphasize access to performance data and to training and technical assistance as incentives can be described using the extended PBAS framework.
- The PBAS framework can be used to describe the evolution of PBAS designs as responses to changing performance problems. This is important to the expanded learning field, which is moving from lower to higher-stakes designs.

Findings for Reliability of Program Ratings and Validity of Performance Levels

While developmental science may eventually describe critical thresholds of youth experience, above or below which effects on individual learning and development will occur, we do not currently have such evidence. We reviewed literature from the early childhood field, which provides exemplars for methods and some clues about quality thresholds for younger children. Among our findings:

- Extrapolating to the expanded learning field, it is unlikely that the high level of quality above which effects diminish can be identified, or that our measures are constructed to detect such a level. For now, a safe bet is that higher quality is better, although the range of measurement tools available probably do not define a very high level of quality.
- Extrapolating to the expanded learning field, it is more likely that low quality has harmful developmental effects. Our priority should be to closely examine the relationship between lower quality and social and emotional skills.

The Palm Beach County PBAS has yielded a unique longitudinal data set of program ratings for quality of instructional practices. This data file has allowed us to examine a number of critical questions

related to the reliability of program ratings for instructional quality and validity of performance levels that are set by the PBAS. Findings for reliability of program ratings include:

- The Palm Beach County PBAS composes a program rating for instructional quality as follows: In each program, three offerings are sampled and rated with PBC-PQA Form A. An instructional total score is produced for each offering session by taking the mean across 13 scales. The three mean ratings are then averaged to produce a program rating.
- The composition formula described above produces a program rating above the margin for acceptable reliability for social science and for observational measures used in the field.[34] The rating demonstrates consistency between the three sampled offering session scores within a given program, consistency within scales used to create an offering session score, and consistency of program ratings through time. However, caution is advised in higher-stakes applications because the reliability of the program rating is not sufficient to differentiate between programs of similar quality. Note that in the Palm Beach County PBAS, performance levels are defined for at least two ratings – one for quality management practice and one for quality instructional practices – exemplifying the principle of using multiple measures to make up for potential error in any single measure.
- The program rating scale, ranging between a lowest program score of 2.64 in year one and highest program score of 4.87 in year five, does not appear to have significant ceiling effects. This means that most programs have room to score higher on the scale. However, we again caution that this does not mean that the measure is designed to capture a very high level of quality – a validity issue that the field needs to address.
- Offering type is a potential source of bias in ratings. PBAS for expanded learning should probably include guidance for selection of offering types to be rated. In more sophisticated treatments, where program ratings are predicted using some kind of a measurement model, activity type should be treated as a covariate likely to affect both the relationship between scales and total score, and between the total score and other variables (Bollen & Bauldry, 2011).

[34] In terms of the criteria for evidence of reliability created by Marybeth Shinn for the W.T. Grant-funded compendium of quality measures, the reliability of a Palm Beach County program rating could be described as "strong by general standards" (Yohalem et al., 2009). Their criteria for inter-rater reliability describe an intraclass correlation with "…values close to or above .5 to indicate high reliability" (p. 87). The intraclass correlation for individual offering session scores within a program rating in this study was .44. Similarly, the intraclass coefficients for internal consistency for a Palm Beach County offering session rating were .85, well above the .70 cut off identified by Shinn and colleagues (p. 88). While no quantitative criteria were given for test-retest reliability the authors suggest that, over time, correlations should moderate to reflect stability but not too high to reflect sensitivity to change (p .15). The average, error-adjusted correlation coefficient for any two successive years in the Palm Beach County data file was .66.

Our analyses with the program ratings data file provided an opportunity to examine both the point-in-time distribution of program ratings for each year and the within-program rating trajectories over five years. We were able to ask and answer questions about the validity of the Palm Beach County PBAS more generally, and specifically for performance levels that the PBAS sets for instructional quality ratings. Our findings include:

- The Palm Beach County PBAS performance level for instructional quality is set at a program rating of 4.1 for two successive years, So, importantly, the performance level consists of both a level on the rating scale and a time period during which the level must be maintained. A program has to meet or exceed a rating of 4.1 for two successive years to attain the incentive (public recognition) and drop below the level for two years to lose it.
- As would be predicted by a PBAS theory of change, program ratings appear to go up over time and programs that initially rate high sustain those high quality levels. Mean ratings for all programs in the PBAS over five years increased steadily from 3.71 to 4.16. Trajectories for a sample of low-scoring programs went up steadily over five years, and six of the 10 low-scoring programs were able to attain the high performance level during the period. Initially high-scoring programs stayed high over the five-year period.
- Volatility in program ratings was present year to year, but of the 10 low-scoring and 10 high-scoring programs rated in each year (a total of nearly 100 program ratings), only once did a program that made it over the designated high performance level then dip below that level for more than a single year.
- High and low performance levels can be identified using local performance norms through one of two methods: Programs in the highest and lowest quartiles; or all programs one standard deviation or more above the mean, or one standard deviation or more below the mean.
- Using either method to identify high and low performance bands successfully differentiates between programs by:
    - Identifying subgroups of high and low performance that do not include most of the participating programs, so that incentives are not too costly (i.e., not all of the programs have to be targeted for improvement resources).
    - There is a substantial difference in the magnitude of the scale interval between high and low, increasing confidence that these two subgroups are actually different from each other.
- Using Rasch methods that control for activity type, it is possible to identify performance subgroups (groups of programs that are alike because they tend to cluster in specific parts of the

rating scale). The high and low performance bands produced using these methods are similar to those identified using either the quartiles or standard deviation methods described above.

Recommendations

The following recommendations are offered for Weikart Center clients and others that (a) use the Youth Program Quality Intervention to define a set of quality management practices and (b) use the Youth or School-Age Program Quality Assessment to define a set of quality instructional practices. We provide two types of recommendations. First, we provide a few general recommendations that draw upon findings from the report to address common applied issues of direct and immediate relevance to Weikart Center clients and partners. Second, we provide two sets of detailed recommendations regarding the design and implementation of an expanded learning PBAS. These detailed recommendations should be considered in draft status and will require further treatment, with potential to become a technical manual.

General Recommendations

1) **The Youth PQA Form A can be used to produce a program rating that is sufficiently reliable for linkage to PBAS incentives**. In the Palm Beach County case, three offerings are sampled and rated with the PQA Form A. An instructional total score is produced for each offering session by taking the mean across 15 scales. The three mean ratings are then averaged to produce a program rating. The program rating alone should be linked only to less consequential incentives (e.g. training supports), because the program rating cannot differentiate reliably between programs with ratings that are close on the rating scale. The program rating for quality of instructional practices should be accompanied by other program ratings, such as quality management practices.
2) **PBAS designs should include measures and performance levels for quality management practices with links to incentives for the management role**. Evidence and common sense suggest that management practices – e.g., team self-assessment and planning, real time performance feedback, selection of training focused on specific instructional practices – focused on the quality of service outputs, especially instruction and content, are critical to delivery of high quality services.
3) **PBAS designs should include program ratings for multiple service outputs linked to a mix of incentives.** In the early phases of PBAS implementation, higher-stakes incentives (e.g., funding) should be tied to program attendance and quality management practices, while lower-

stakes incentives should be tied to quality instructional practices and other outputs on the right-hand side of the service production map (see Figure 2). As PBAS participation is integrated into program management, and as performance improves, higher-stakes incentives can be linked to outputs farther to the right of the service production map.

4) **Other best practices for design and implementation of a PBAS might include**:
    - A PBAS design should differentiate staff roles and link performance measures to incentives targeted at specific roles because individual behavior change is the object of a PBAS.
    - A PBAS should emphasize participants' understanding of performance levels and sense of fairness, while evolving toward higher-stakes incentives over time to avoid perverse incentives (e.g., gaming the system).
    - A PBAS should deploy measures to produce multiple program level ratings that are domain specific, i.e., one rating for quality management practices, one rating for quality instructional practices, one rating for youth engagement, etc. There is no evidence and little theoretical justification for the validity of a single combined rating that includes program ratings across multiple domains.

<u>Detailed recommendations related to the Design Standard for Expanded Learning PBAS (Draft)</u>

1) *Goals* - Should be determined by stakeholders and should make clear the ultimate desired outcomes of the service so that customers give the PBAS full consideration and staff give it a full level of effort toward the right goals. These goals might wisely be thought of as a set of value propositions for funders of the expanded learning service.
2) *Service Production Map* – Expanded learning systems should work with program stakeholders to define key service outputs that include both setting processes and individual skill domains. Minimally, setting processes include quality management practices at the organization level and quality instructional practices at the point-of-service level. For programs focused on school success, quality content should be addressed as alignment with school-day content and staff. Individual skill domains should not be included unless programming actually reflects efforts to improve the noted domains.
3) *Measures* – Recommended measures for a PBAS include: major service interruption, endorsement for full participation in PBAS, staff job satisfaction, quality management practices, quality instructional practices and content, youth engagement, and program attendance. Individual skills and beliefs measures can be included if the program design targets specific

skills. Summary recommendations for measures and methods is provided in column one of Table 9.

4) *Ratings and Levels Map* – Measures used in a PBAS need to reflect the program as the focal unit/level of analysis. Care should be given to how measures of lower-level objects (instructional quality, individual skills) are composed into program-level ratings, with specific attention to reliability of the measures for lower-level units (e.g., rater and items). Program ratings for specific output domains (e.g., quality management practices, quality instructional practices, youth engagement) should not be combined into a single rating without evidence for the validity of that composition. Performance levels for program ratings must be clearly defined if performance data is to be linked to incentives. Without independent evidence of validity for specific performance levels,[35] it is probably best to use local norms (high and low quartiles, one standard deviation above or below) to determine performance levels. Summary recommendations for composition of measures and focal performance levels are provided in columns 2 and 3 of Table 9.
5) *Incentives* - These should include a mix of higher and lower stakes incentives, including funding at various levels, review of rating by external customers and internal stakeholders, access to training and technical assistance for improvement methods, and access to performance data.
6) *Performance Improvement Map* – An effective PBAS in the expanded learning field[36] should include supports for performance improvement. These supports could include both methods for performance improvement (e.g., the YPQI) and financial and other supports to learn and implement these methods.
7) *Effective Performance Data* – Performance data should be timely, objective, reliable, sensitive, valid, and feasible. The data should also serve multiple purposes and be drawn from and seek to develop and improve performance at multiple levels of an organization. These characteristics are defined in Appendix B.

[35] For example, validity evidence is available for performance levels for quality management practices based on the YPQI method and for instructional practices measured by the Youth PQA and focused on youth engagement.
[36] As we argued in the introduction, the strengths of the expanded learning field include many unmeasured qualities of community responsiveness and cultural resources. A more pure market-based approach may select vendors who can excel on PBAS measures but who lack these unmeasured qualities. Further, the idea that PBAS designs are intended to actually improve all performances rather than just eliminate weak performers is a stated goal in most expanded learning systems.

Table 9. Summary Recommendations for Rating Composition and Performance Levels

Measure	Method; Source	Composition of Rating
Program Instability	Criteria set in advance, e.g. manager turnover; Annually updated records	Y/N
PBAS Participation Fidelity	Certification of training status for program staff; annually updated records	Y/N
Job Satisfaction	Five items all staff; annual survey	Mean items individual score; mean over staff for program rating with flag for programs with high disagreement
Quality Management Practices	15 items all staff; manager interview PQA Form B	Mean items individual score; mean over staff for program rating with flag for programs with high disagreement
Quality Instructional Practices	PQA Form A, safety items plus 13 scales; external observer or guided program self-assessment	Mean scales for offering session total score; ,ean offering session scores for program rating; self-assessment for lower stakes
Quality of Content/ School Alignment	12 items all staff; annual survey	Mean items individual score; mean over staff for program rating with flag for programs with high disagreement
Engagement	8 items all youth; annual survey	Mean for a sample of offering sessions
Attendance	TBD	Determined by funder
Youth skills/beliefs	Surveys, test data	Change score

Detailed Recommendations related to PBAS Implementation Phases (Draft)

Application of the PBAS nomenclature and framework to describe expanded learning systems has produced capacity to describe the evolution of the Palm Beach County PBAS toward multiple program ratings with links to incentives, and toward a PBAS design that includes higher-stakes elements. In this section we draw upon the Palm Beach County case study but add measures that are used in other expanded learning PBAS', as summarized in Appendix A. In Table 10 below, we describe a hypothetical PBAS design that links measures to performance levels and incentives. Nine measures are included in the design, each described in Table 9. Four types of incentives are included in the design: funding for operations or incentives (F), customer review by making the program ratings public (C), supervisory review for low performance (S), and access to data for program improvement (D). Other types of incentives are available. Lowercase designators for the high performance group (hi) and the low performance group (lo) describe the focal performance level for the measure that triggers the incentive. The three phases include a first phase focused on adopting measures and identifying performance levels

on the rating scales. In this phase, PBAS participants are exposed to the measures and data, and local norms for high and low performance can be reviewed. In the second phase, incentives are attached to specific performance levels on various measures. In subsequent phases, these steps are repeated for outputs further to the right-hand side of the service production map (see Figure 2) and adjustments to prior performance levels and incentives can be made.

Several more specific aspects of Table 10 are worthy of note. First, access to data is an incentive that exists for all measures at all levels of performance in all phases – if the data is suited for use as part of an improvement method. Second, some performance levels, like program stability and attendance, are important at all times. For example, low performance on these measures would presumably warrant supervisory review during all phases. Third, funding is the potentially highest stake incentive and is linked to compliance with all PBAS requirements, implementation of quality management practices and youth attendance. In Table 10's Phase 2 column, supervisory review is added where scores indicate low staff job satisfaction, low quality of instruction, and low levels of youth engagement. Incentives related to public review are high compliance with PBAS requirements and high ratings for the quality of services – primarily because these are the markers of high quality service that can be easily summarized for consumption by customers. Finally, growth in youth skills is added in a subsequent phase because this is by far the most difficult measure to implement and for which to identify performance levels. The performance level for individual skills and beliefs is an increment of growth in a particular skill domain, which introduces a number of technical challenges for which measures in the field are ill prepared to address.

Table 10. Phases for Alignment of Levels with Incentives

	Phase 1: Align measures and set levels	Phase 2: Add incentives	Subsequent Phases: Repeat and adjust
Program Instability	**S-lo**		
PBAS Participation Fidelity	**C-hi**	**F-hi**	
Job Satisfaction	**D**	**S-lo**	
Quality Management Practices	**D**	**F-hi**	
Quality Instructional Practices	**D**	**S –lo, C-hi**	
Quality of Content/ School Alignment	**D**		
Engagement	**D**	**S -lo**	
Attendance	**D, S-lo, F-hi**		
Youth skills/beliefs	**D**		**S-lo**

Incentive Types: F=funding; C=customer review; S=supervisory review; D=access to data
Focal Performance Level: Hi= focus on "good" perf. level; Lo=focus on "bad" perf. level

References

Aguinis, H, Joo, H, & Gotfredson, R. (2012). Performance management universals: Think globally, act locally. *Business Horizons, 55*, 385-392.

Akiva, T., Smith, C., Sugar, S., & Brummet, Q. (2011). *Staff instructional practices, youth engagement, and belonging in out-of-school time programs*. Paper presented at the American Educational Research Association Annual Meeting.

Arkansas Department of Human Services. (2013). Better Beginnings: Every Child Deserves our Best. Retrieved July 15, 2013

Baker, M., Gruber, J., & Milligan, K. (2008). Universal Child Care, Maternal Labor Supply, and Family Well-Being. *Journal of Political Economy, 116*(4), 709-745. doi: 10.1086/591908

Bliese, Paul. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. Klein & S. Kozlowski (Eds.), *Multi-level theory, research and methods in organizations* (pp. 349-381). San Francisco: Josey-Bass.

Bohnert, A., Fredricks, J., & Randall, E. (2010). Dimensions of Youth Organized Activity Involvement: Theoretical and Methodological Considerations. *Review of Educational Research, 80*(4), 576-610.

Bollen, Kenneth A., & Bauldry, Shawn. (2011). Three C's in measurement models: Causal indicators, composite indicators and covariates. *Psychological Methods, 16*(3), 265-284.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Education Measurement, 32*(4), 385-396.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology, Vol. 1. Theoretical models of human development (6th Ed.)* (pp. 793-828). New York: Wiley.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early childhood research quarterly, 25*, 166-176.

Camm, F., & Stecher, B. (2010). Analyzing the the operation of performance-based accountability systems for public services. Santa Monica: Rand Corporation.

Collaborative for Building After-School Systems. (2007). Shaping the future of after-school. In T. A. S. Corportation (Ed.), *The essential role of intermediaries in brining quality after-school systems to scale* (pp. 1-20). New York: The collaborative for Building After-School Systems.

David P. Weikart Center for Youth Program Quality. (2012). Youth Program Quality Intervention Report: 2012 Findings from the Washington State 21st CCLC Program: Report to the Washington Office of Superintendent of Public Instruction (OSPI) *YPQI Report*. Ypsilanti, Michigan: David P. Weikart Center for Youth Program Quality.

David P. Weikart Center for Youth Program Quality. (2013a). Nashville After Zone Alliance Quality Improvement Intervention: 2012-2013 Findings from the Northeast, South Central, and Northwest Zones. Ypsilanti, Michigan.

David P. Weikart Center for Youth Program Quality. (2013b). United Way of Greater Kansas City Out-Of-School Time Quality Matters Project.

Diamantopoulos, Adamantios. (2008). Formative indicators: Introduction to the special issue. *Journal of Business Research, 61*, 1201-1202.

Dodge, K. A. (2011). Context matters in child and family policy. *Child Development, 82*(1), 433-442.

Durlak, J. A., Weissberg, R. P., & Pachan, M. K. (2010). A Meta-Analysis of After-School Programs That Seek to Promote Personal and Social Skills in Children and Adolescents. *American Journal Community Psychology*, 16.

Early, D.et al. (2007). Teacher education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development*.

Eccles, J., & Gootman, J. (Eds.). (2002). *Community programs to promote youth development* Washington, DC: National Academy Press.

Fischer, K.W., & Bidell, T. R. (2006). *Dynamic development of action, thought, and emotion* (6th ed. Vol. 1). New York: Wiley.

Gambone, M. A., Klem, A. M., & Connel, J. P. (2002). Finding out what matters for youth: Testing key links in a community action framework for youth development. Philadelphia: Youth Development Strategies Inc. & Institute for Research and Reform in Education.

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. . (2007). The new instructional leadership: Creating data-driven instructional systems in schools. Madison, WI: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Harvey, Robert J., & Hollander, Eran. (2004). *Benchmarking rwg interrater agreement indices: Let's drop the .70 rule-of-thumb*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Haut, Sheryl R.et al. (2002). Interrater reliability among epilepsy centers: multicenter study of epilepsy surgery. *Epilepsia, 43*(11), 1396-1401.

Honig, M. (2004). The new middle management: Intermediary organizations in educational policy implementation. *Educational Evaluation and Policy Analysis, 26*(1), 65-87.

Howes, C., Phillips, D., & Whitebook, M. . (1992). Thresholds of quality: Implications for the social development of children in center-based child care. *Child Development, 63*, 449-460.

James-Burdumy, S.et al. (2005). When schools stay open late: The national evaluation of the 21st century community learning centers program final report *Mathematica Policy Research, Inc.*

James, Lawrence R., Demaree, Robert G., & Wolf, Gerrit. (1993). rwg: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306-309.

Kania, J., & Kramer, M. (2011). Collective impact. *Stanf Soc Innov Rev. Winter*, 36-41.

Kim, Sungsook C, & Wilson, Mark. (2008). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of applied measurement, 10*(4), 408-423.

Knockaert, M., & Spithoven, A. (2012). Technology intermediaries in low tech sectors: the case of collective research centres in Belgium. *Innovation: Management, Policy, & Practice, 14*.

La Paro, Karen M., Pianta, Robert C., & Stuhlman, Megan. (2004). Classroom assessment scoring system (CLASS): Findings from the prekindergarten year. *The Elementary School Journal, 104*(5), 409-425.

Larson, R., & Angus, R. M. (2011). Adolescents' development of skills for agency in youth programs: Learning to think strategically. *Child Development, 82*, 277-294.

Larson, R., & Brown, J. R. . (2007). Emotional development in adolescence: What can be learned from a high school theater program? *Child Development, 78*(4), 1083-1099.

Larson, R., Hansen, D., & Moneta, G. (2006). Differing profiles of developmental experiences across types of organized youth activities. *Developmental Psychology, 42*, 849-863.

Linacre, J.M. (2004). Optimizing rating scale category effectiveness. In J. Smith, E.V. & R. M. Smith (Eds.), *An introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Lowenstein, A. (2011). Early care and education as educational panacea: What do we really know about its effectiveness. *Educational Policy, 25*(1), 92-114.

Mashburn, A., & Pianta, R. (2010). Opportunity in early education: Improving teacher-child interactions and child outcomes. In A. Reynolds, A. Rolnick & J. Temple (Eds.), *Cost effective programs in children's first decade: A human capital integration*. New York: Cambridge University Press.

Mashburn, A.et al. (2008). Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development, 79*(3), 732-749.

Mason, S.A. (2003). *Learning from data: The roles of professional learning communities*. Paper presented at the American Educational Research Association, Madison, WI. http://www.eric.ed.gov/PDFS/ED476852.pdf

McCartney, Kathleenet al. (2010). Testing a series of causal propositions relating time in child care to children's externalizing behavior. *Developmental Psychology, 46*(1), 1-17. doi: 10.1037/a0017886

Naftzger, N. (2012). Using Rasch Modeling Techniques to Address Issues of Reliability and Validity in Observational Measures: A Case Study Employing the School-Age Youth Program Quality Assessment. In A. I. f. Research (Ed.): William T. Grant Foundation.

Naftzger, N.et al. (2012). Texas 21st Century Community Learning Centers: Interim Report (pp. 127). Naperville, IL: American Institutes for Research.

National Child Care Information and Technical Assistance Center. (2009). Quality Rating Systems: Definition and Statewide Systems. Fairfax, VA: National Child Care Information and Technical Assistance Center.

National Research Council and Institute of Medicine. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities* (M. Oconnel, T. Boat & K. Warner Eds.). Washington DC: Board on Chiildren, Youth and Families, Division of Behavioral and Ssocial Sciences, National Academies Press.

Nunnally, J. C. (1978). *Psychometric theory (2nd ed.)*. New York: McGraw-Hill.

Oyserman, Daphna, Bybee, Deborah, & Terry, Kathy (2006). [Possible Selves and Academic Outcomes: How and When Possible Selves Impel Action].

Rivard, K. (2012). Out-of-School Time Quality Matters Participation Bonus Award. In U. W. o. G. K. City (Ed.).

Schweig, J. (2013). Measurement error in multi-level models of school and classroom environments: Implications for reliability, precision and prediction (pp. 1-40): National Center for Reserach on Evaluation, Standards and Student Testing.

Smith, C., Akiva, T, Gersh, A., & Sutter, A. (2012). Feasibility Study for Impact Evaluation and Intervention Design Improvements: Public Report (pp. 104): David P. Weikart Center for Youth Program Quality, a division of the Forum for Youth Investment and Prime Time, Inc.

Smith, C.et al. (2012). *Continuous quality improvement in afterschool settings: Impact findings from the Youth Program Quality Intervention study* (F. f. Y. Investment Ed.). Ypsilanti, MI: Forum for Youth Investment.

Smith, C.et al. (2012). Development and early validation evidence for an observational measure of high quality instructional practice for science, technology, engineering and mathematics in out-of-school time settings: The STEM supplement to the Youth Program Quality Assessment (pp. 1-25). Ypsilanti, MI: The David P. Weikart Center for Youth Program Quality, a division of the Forum for Youth Investment and Providence Afterschool Alliance.

Smith, Charles, Akiva, T., Blazevski, J., Pelle, L., & Devaney, T. (2008). Final report on the Palm Beach Quality Improvement System pilot: Model implementation and program quality improvement in 38 after-school programs. Ypsilanti, MI: High/Scope Educational Research Foundation.

Smith, Charles, Akiva, T., Devaney, T., & Sugar, S. (2009). Quality and accountablility in the out-of-school time sector. In R. Granger, K. Pittman & N. Yohalem (Eds.), *New Directions for Youth Development: Defining and measuring quality in youth programs and classrooms* (Vol. 121). San Francisco: Jossey-Bass.

Smith, Charles, & Akiva, Tom. (2008). Quality accountability: Improving fidelity of broad developmentally focused interventions. In H. Yoshikawa & B. Shinn (Eds.), *Transforming Social Settings: Towards Positive Youth Development*: Oxford University Press.

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

Smith, Charles, Peck, Stephen J., Denault, A., Blazevski, J., & Akiva, Tom. (2010). Quality at the point of service: Profiles of practice in afterschool settings. *American Journal of Community Psychology, 45*, 358-369.

Spielberger, J., & Lockaby, T. (2006). The Prime Time Initiative of Palm Beach County, Florida - QIS development process evaluation: Year 2 report. Chicago: Chapin Hall Center for Children at the University of Chicago.

Spielberger, J., & Lockaby, T. (2008). Palm Beach County's Prime Time Initiative: Improving the quality of after-school programs. Chicago: Chapin Hall Center for Children at the University of Chicago.

Spielberger, J., Lockaby, T., Mayers, L., & Guterman, K. . (2009). *Ready for prime time: The first year of implementation of an afterschool quality improvement system by prime time beach county, inc.*, University of Chicago, Chicago.

State of Vermont. (2013). Step Ahead Recognition Systems (STARS). Retrieved July 15, 2013

Stecher, B.et al. (2010). Toward a culture of consequences: Performance-Based accountability systems for public services. Santa Monica: Rand Corporation.

Sugar, S., Pearson, L., Devaney, T., & Smith, C. (2009). Findings from year 2 of the Palm Beach County quality improvement system, 2008-09: Quality standards in 90 afterschool programs. Ypsilanti, MI: The David P. Weikart Center for Youth Program Quality.

Wang, Lihshing. (2010). Dissatenuation of correlations with fallible measures. *Newborn and infant nursing review*, 60-65.

Weitzman, B.C., Silver, D., & Brazill, C. (2003). Efforts to Improve Public Policy and Programs Through Improved Data Practice: Experiences in Fifteen Distressed American Cities. New York University: Robert Wood Johnson Foundation.

Wilson-Ahlstrom, A., Yohalem, N., DuBois, D., & Ji, P. (2011). From Soft Skills to Hard Data: Measuring youth program outcomes. In T. F. f. Y. Investment (Ed.). Washington, DC: The Forum for Youth Investment.

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 and The Wallace Foundation.

Yohalem, N.et al. (2010). Making quality count: Lessons learned from the Ready by 21 Quality Counts Initiative. Washington, D.C.: The Forum for Youth Investment.

Yohalem, N., Ravindranath, N., Pittman, K., & Evennou, D. (2010). Insulating the education pipeline to increase postsecondary success. Washington, DC: Forum for Youth Investment.

Yohalem, N., Wilson-Ahlstrom, A., Fischer, S., & Shinn, M. (2009). Measuring youth program quality: A guide to assessment tool, second edition. Washington, DC: The Forum for Youth Investment.

Zaslow, M.et al. (2010). Quality Dosage, Thresholds, and Features in Early Childhood Settings: A Review of the Literature, OPRE 2011-5. *Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services*.

Zaslow, M., Tout, K. , Halle, T., & Forry, N. (2009). Multiple Purpose for Measuring Quality in Early Childhood Settings: Implications for Collecting and Communicating Information on Quality. *OPRE Issue Brief,* (2), 1-11.

Zellman, G. , Perlman, M., Le, V., & Setodji, C. M. . (2008). Assessing the validity of the qualistar early learning quality rating and improvement system as a tool for improving child-care quality. In R. Corporation (Ed.), (pp. 1-130). Santa Monica, CA: RAND Corporation.

Appendix A – Expanded Learning PBAS in Kansas City, Oakland, Vermont and Arkansas

Kansas City

The United Way of Greater Kansas City's expanded learning PBAS began in 2009 with a pilot and has since grown to 120 program sites. Table A-1 describes key components. Broad, publically stated goals for the initiative include seeking to "strengthen child and youth achievement, help youth overcome barriers to success, and maximize long-lasting benefits" for "at-risk, low-income students." Key outputs to achieve these goals include "effective programs with staff who engage young people in learning during non-school hours" with the Youth and School-Age Program Quality Assessments as the core measures for program ratings. The PBAS includes performance improvement supports and methods: a high capacity quality intermediary organization (QIO; Francis Institute) supports the YPQI and provides on-site coaching. Primary incentives include funding, access to QIS supports, and network recognition.

The United Way of Greater Kansas City offers YPQI participation bonus awards to sites that implemented the YPQI process based on three defined levels of fidelity (see table below). At the end of the 2012-2013 program year, United Way of Greater Kansas City recognized 39 sites for high levels of implementation to the YPQI process. This number represents an increase of 13 sites over the 2011-2012 program year.

Table A-1: Key Components of Kansas City PBAS

Goals	**Measures**	**Incentives**
Quality management practices High quality instruction by staff Child/youth development and learning	Count of YPQI fidelity Program rating (Youth and School-Age PQA)	Financial bonus based on performance levels Access to performance supports and methods Network recognition

Definition of levels	**Levels mapped to incentives**	
Participation Level 1: Manager completes the following by the deadlines: (1) team self-assessment data entered on *Online Scores Reporter*; **(2)** Program Improvement Plan entered on *Online Scores Reporter*; **(3)** 40% goals met; **(4)** Attend three skills trainings with 1-4 site staff may attend. **Participation Level 2**: Same as level 1 but 60% of goals met and 5 trainings attended. **Participation Level 3**: Same as level 1 but 75% of goals met and 7 trainings attended	Level 1:	$300
	Level 2:	$500
	Level 3:	$750

Citations: (Rivard, 2012)

Oakland

The Oakland Out-of-School Time (OST) PBAS is conducted under the auspices of evaluation by Public Profit Inc. and encompasses 92 programs funded by the Oakland Unified School District (OUSD) and 67 programs under the responsibility of the Oakland Fund for Children and Youth. Table A-2 describes key components of the PBAS. The goals of OUSD's afterschool programs are stated by

legislation as being designed to provide children and youth with safe and educationally enriching alternatives during nonschool hours, including literacy, academic enrichment, and safe constructive alternatives. The Oakland OST evaluation has been modeled on the YPQI since 2010, with a few modifications. It is optional for sites to conduct a program self-assessment. However, all sites are observed twice per year for an external assessment. All sites receive a detailed report that includes: Youth PQA data, survey and interview information, and a narrative summary of the program strengths and areas for improvement. Sites also have access to Youth Work Methods Workshops throughout the cycle in order to build staff skills in positive youth development practices.

Table A-2: Key Components of Oakland PBAS

Goals	Measures	Incentives
Academic achievement High quality program services Keep city council and school district informed about performance	Program rating on Youth PQA and PQA Academic Supplement Youth participation records Academic records	Performance feedback Public recognition

Definition of levels	**Levels mapped to incentives**
• **Emerging** – Program is not yet providing high-quality service. Defined as a site that has three or more domains with 25% or more "1" ratings. • **Performing** – Program assures participants' physical and emotional safety (defined as having less than 25% "1" ratings in Safe and Supportive), and has a few areas for additional improvement. Defined as a site with up to two domains with 25% or more "1" ratings in Interaction, Engagement, or Academic Climate. • **Thriving** – Program provides high-quality services across all five quality domains. Defined as a site with no domains with 25% or more "1" ratings.	Performance levels are published annually in the system evaluation report

Vermont

The Vermont Center for Afterschool Excellence (now Vermont Afterschool, Inc.) PBAS began in 2010 with a pilot and now includes 25 sites. Table A-3 describes key components. The Vermont PBAS is designed to: (a) build program leaders' continuous quality improvement skills; (b) increase the quality of instructional practices delivered in afterschool programs; and, ultimately, (c) increase students' engagement with program content and opportunities for skill-building. The Vermont Center for Afterschool Excellence uses the YPQI model and develops site mentors. In 2012-2013, 21st Century Community Learning Center sites were mandated to participate, the Youth PQA was listed as an accepted tool by the state QRIS, and programs participating in the 2012-2013 academic year were provided with external site assessments. The external assessment scores were tied to particular point levels in the Step Ahead Recognition System (STARS).

STARS programs receive higher ratings when they "go above and beyond state regulations to provide professional services that meet the needs of children and families." They also receive a higher rate of funding from the Child Care Financial Assistance Program, which "pays a higher rate on behalf of families … based on the number of stars the program has earned." Primary incentives include community awards (discounts at local businesses), funding, and local and statewide recognition. Other incentives include community awards and financial rewards (e.g., higher reimbursement on the child care financial assistance fee scale).

Table A-3: Key Components of Vermont PBAS

Goals	**Measures**	**Incentives**
• Build program leaders' quality improvement skills • Increase the quality of instructional practice • Increase student engagement	Providers may apply for STARS recognition in five areas: • Compliance with state regulations • Staff qualifications and training • Interaction with and support of child, family, community • Quality of assessment and planning for improvements • Quality of operating policies and business practices.	Funding • Reimbursement of the child care financial assistance fee • Opportunity to apply for grants open to nationally accredited programs • Discount on purchases from national/Vermont companies. Public recognition(if requested) • listing on the STARS website • supply of STARS brochures • customized press release

Definition of levels	**Levels mapped to incentives**
One-star programs (1-4 points) examining practices to enhance services. They may be fairly new, just starting on a path of improvement and growth, or be stronger in one area. **Two-star programs (5-8 points)** making commitment to strengthen practices. Have made progress in many areas or more progress in one or two areas. **Three-star programs (9-11 points)** have made improvements and working toward goals. Have made substantial progress in two or three areas or made some improvements in all five areas. **Four-star programs (12-14 points)** are established programs that have met several standards of quality in all five areas. Many four-star programs are also nationally accredited. **Five-star programs (15-17 points)** are outstanding in all five areas. Many five-star programs are also nationally accredited.	The Child Care Financial Assistance Program (CCFAP) pays a higher rate based on the number of stars earned. • 1 Star – 5% above base CCFAP rate • 2 Stars – 10% above base CCFAP rate • 3 Stars – 20% above base CCFAP rate • 4 Stars – 30% above base CCFAP rate • 5 Stars – 40% above base CCFAP rate Bonus payments for EACH level achieved: Star - $250; 2 Stars - $500; 3 Stars - $1,000; 4 Stars - $1,150; 5 Stars - $1,550
Citations:(State of Vermont, 2013)	

Arkansas

The Better Beginnings Program has been Arkansas' PBAS since 2010, a voluntary system for afterschool care providers to measure continuing progress in areas of staff training, program quality and facility standards. Better Beginnings is a multistep certification process that begins with completion of the Better Beginnings Application Checklist. Providers can use this self-assessment checklist to identify requirements they are already meeting and those additional requirements they will need to meet in order

to receive certification. Through the Better Beginnings website (http://www.arbetterbeginnings.com), providers can also access child care provider tool kits that will help them begin addressing those areas they need to improve. Participating programs' quality ratings are published on the Better Beginnings website. Scores above the level of "1" require reaching a threshold of quality on the Youth or School-Age PQA. Incentive grants allow an administrator to make decisions and use the grant funds to support increasing and/or maintaining the quality components of the facility as part of the plan for improvement.

Table A-4: Key Components of Arkansas PBAS

Goals	Measures	Incentives
Provide a resource for parents and guardians looking for information about high quality childcare. Provide incentives for child care providers to participate in a quality improvement system.	Program rating on School-Age PQA, Youth PQA or Environmental Rating Scale On-site review of Better Beginnings requirements, Program and Business Admin Scales.	• Public recognition – rating published • Incentive grants if meet certification standards at levels 1, 2, and 3. Amounts: http://www.arbetterbeginnings.com/wp-content/uploads/2011/06/BBIncentiveAmts.pdf
Definition of levels		**Levels mapped to incentives**
Level 1: Manager trained on self-assessment. **Level 2**: Programs rating 3.00 or higher on PQA **Level 3**: Programs rating 3.75 or higher on PQA All levels also have additional requirements around professional development, staff certification and curriculum.		Public recognition Incentive grants renewable for levels 1 and 2 and annually available for level 3. Amounts based on licensed capacities, current and prior performance
Citations: (Arkansas Department of Human Services, 2013)		

Appendix B: Optimal Characteristics of Performance Data

In order to effectively use performance measures and data to impact performance goals, the data need to meet certain requirements. This brief describes the optimal characteristics for performance data so that they can be acted upon for improvement. Eight important characteristics are that data should be timely, objective, reliable, sensitive, valid, feasible, multi-purpose, and multi-level.

Timely. Data that are available in real time as events occur or just after completion are more likely to hold relevance for actors. While historical data and trends can be informative, if they do not hold personal significance for the actor, they are not likely to impact performance. For example, a teacher reviewing her students' quiz results from Friday can use that information to change her instructional focus for Monday. The quiz results will not be as useful even two weeks from now when the class has moved on to a different topic.

*Objective***.** Objective data are focused on behaviors and conditions that can be identified through observation and easily named in relation to practice. Objectivity helps to provide concrete feedback to actors, and minimize subjective feelings of judgment. Observations provide information about what actually occurred, allowing the actor to consider the underlying intentions and causes for the behavior or condition. For example, calling to attention the number of times an instructor calls on boys as opposed to girls could start a discussion about gender equity in the program.

Reliable. Data should be seen by all stakeholders as precise and factual due to standardization of measures/methods. Reliability of an instrument is increased by identifying the precise data needed and through repeated use of the instrument in field testing. Reliable data yields similar results after similar analyses. For example, a classroom observation checklist – if it contains criteria that are valued by all stakeholders and has a clear methodology for its use – can be used to collect reliable data.

Sensitive. Optimal performance data describe behaviors and conditions that are likely to change in response to intervention. Optimal performance measures capture that change. Sensitivity can be a difficult characteristic to attain for some data. For example, detecting change in skill development can be challenging due to the inconsistent nature of skill development and potential for learners to regress one or more times during the process of attaining mastery. For more linear and predictable data, sensitivity may be easier to accomplish (e.g., measuring the height of a child over time with a tape measure). Still, the length and intensity of the intervention as well as the measure used can impact the sensitivity of the data. You are not likely to see a change in height of a child in a two-week time period, or if you are measuring in full inches rather than in fractions of an inch.

Valid. Data are valid when they describe behaviors and conditions thought to be a link in a causal chain of events desired by the actors involved. This is called *constructive* validity – the reasoning and evidence behind the argument follows a logical trajectory. Data can also have *substantive* validity – an

argument whose conclusions are of great value to the stakeholders or to a field of research. For example, for valid instructional performance data, there is an answer (from research, consultation with experts, or consultation with practitioners) to the question, "Which instructional practices are important and where can an observer see them?"

Feasible. Data collection must be a feasible process; the minimum data necessary are collected using typical community resources and by typical survey respondents. If data collection goes beyond the means of the organization or causes major disruption to normal operations, it is not feasible.

Multipurpose. When *both* data collection and data interpretation processes promote learning and coordination among actors in the organization, the performance data are truly multi-purpose. A well-engineered process for data collection and interpretation can create a shared language among actors and a framework to guide discussions about performance. For example, a professional learning community of instructors might first observe each other using a rubric of behaviors deemed as best practices. Then they might discuss their observations and score the rubric together, amid discussions of their own practice and how it might be improved.

Multilevel. Data designed for use by individual units (e.g., staff or sites) can be made useful at other levels (e.g., organizations or managers) when aggregated across individual units to assess collective performance. Policies regarding accountability and confidentiality are important for waylaying fears actors may have that the data be used for penalization rather than performance improvement. Purposes of data use at the various levels may also be slightly different; while individual staff or sites might be interested in improving the services they provide, managers or organizations might be interested in using the data for decisions about how best to allocate resources among program sites.

Appendix C: A Generic Logic Model of Expanded Learning Contexts and Skills

This brief describes a model that can be used to map program performance goals to the associated inputs, process, outputs, and outcomes. The model, once specified for the program, can be used to guide measurement selection and evaluation design.

Background

Out-of-school time (OST) programs can provide important spaces for positive youth development that, as a young person develops, complement their growth in school, home, and other contexts (Yohalem, Ravindranath, Pittman, & Evennou, 2010). Sustained OST exposure in particular afterschool settings has been associated with positive change in numerous measures of beliefs, goals, and skills: positive self-perceptions, bonding to school, and positive social behavior (Durlak, Weissberg, & Pachan, 2010); emotional regulation (Larson & Brown, 2007); 'balanced' possible selves (Oyserman, Bybee, & Terry, 2006); initiative and strategic thinking (Larson & Angus, 2011; Larson, Hansen, & Moneta, 2006). With a few caveats, higher levels of participation (intensity, duration, and breadth) are associated with many of these types of outcomes (Bohnert, Fredricks, & Randall, 2010). However, the specification this research provides about how contexts actually produce individual change, or how individual development emerges through time in and across settings, can be difficult to generalize into specific practices and curriculum.

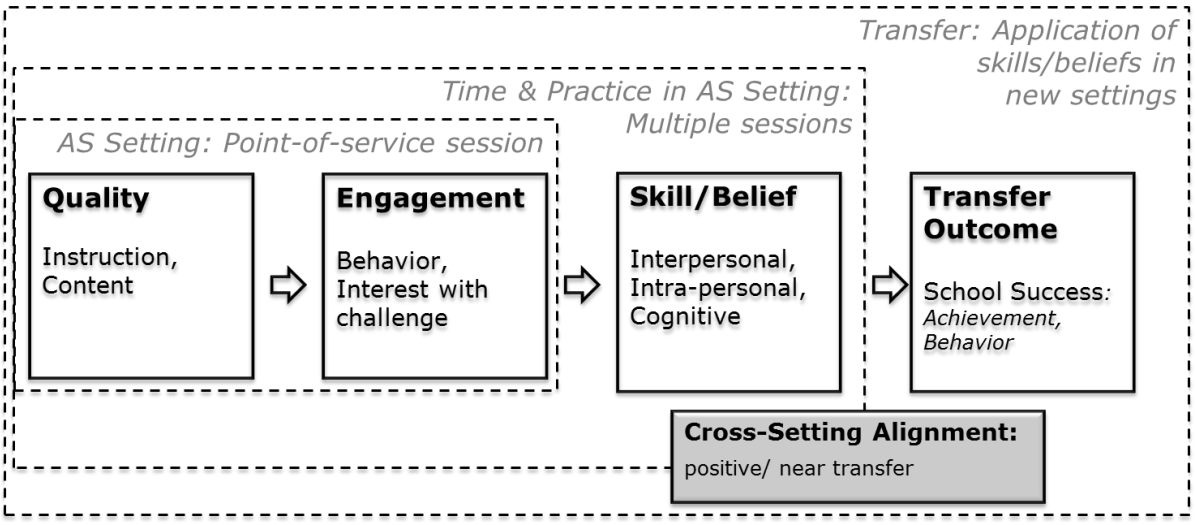
Furthermore, most skills are designed to perform specific functions in particular settings. Transfer of skills from one context to another is a challenge. According to educational researcher Kurt Fischer, context can include the environment or setting where the skill is being carried out; the range of emotional and biological states that occur within a person; and the relationship to other people and the levels of support, challenge, or stress that they provide (Fischer & Bidell, 2006). For example, a student who learns to use the scientific process in biology class might not realize that the same pattern of thinking can be used to deconstruct one author's argument for the onset of the Civil War in history class. Likewise, talented baseball players may not be able to transfer their throwing skills to the football field. When a learner experiences repeated modeling and scaffolding of skills within a positive youth development context, that person is more likely to develop those skills. However, until the learner is able to generalize his or her knowledge and skills (after expansive and repetitive guided practice), the skills are dependent on the context in which they are learned. This challenge is potent in afterschool programs that wish for youth to transfer skills learned in the afterschool program to other contexts, including school, family and employment.

Theory of Change

Figure 1 presents a likely pathway for youth development and learning in OST settings, and for the eventual transfer of skills from the OST context to other contexts (e.g., school success). Figure 1 is intended to support practitioners to consider individual change in ways that (1) support intentionality in

program planning and delivery and (2) make more efficient use of resources committed to measurement, evaluation and continuous improvement.[37]

Figure C1: Theory of OST Contexts and Child-level Change



The primary chain of effects described in the QuEST (Quality, Engagement, Skills, Transfer) model shown in Figure C1 suggests that the quality of instruction and content, delivered at the point of service where teachers and students meet, will produce heightened levels of student engagement with content during afterschool offerings. Over multiple afterschool offering sessions (time) the combination of high quality instruction, content and student engagement, will result in the emergence of beliefs and skills. With sufficient intensity of exposure to high quality environments, specific skills and beliefs will transfer to other settings, including school day classrooms.

*AS Setting: Point-of-service session***.** According to Figure C1, high quality instruction and content produces youth engagement during a given session. The point-of-service setting is the place where staff, youth, and resources come together as activities (Smith & Akiva, 2008; Smith et al., 2010) and is a youth-in-context transactive system (cf. "microsystem" in Bronfenbrenner & Morris, 2006). That is, youth bring their experiences, background, motivation, attitudes, etc., to the point of service, and the setting provides features that include instructional practices and content. During each afterschool session, qualities of experience are provided by the context (quality of instruction and content) and produce youth engagement.

[37] The confusing array of names for measures of child and youth outcomes is part of the reason for creating Figure 1, which is designed to help practitioners think clearly about the outcomes they are trying to achieve and empower them to review actual item content, rather than more abstract scale names.

Time and Practice in AS Setting: Multiple sessions. The simultaneous presence of high quality instruction and high youth engagement across multiple sessions produces mastery experiences related to the process and content of the sessions. Youth engagement over multiple sessions is likely to include regular experience of positive effect, concentration on tasks requiring moderately-difficult effort, and receipt of scaffolding – especially adults' modeling of the learning task (Fischer & Bidell, 2006) which can be socioemotional (e.g., using your words), academic (e.g., reducing fractions), or expressive (e.g., design a service project). A sequence of high quality, high engagement sessions over time leads to development of specific skills ("practice, practice, practice"). This is the point where the theory of change requires program providers to think specifically about which skills they are trying to grow and how likely the qualities of program experience they provide will grow those skills.

Transfer: Application of Skills/Beliefs in New Setting. The third box raises the issue of skills transfer and the likelihood that the skills are mastered well enough to be applied in other settings. Context-specific mastery experiences support longer-term skill development and skill transfer to external settings, leading ultimately to improved outcomes of interest to policymakers. In our model, youth engagement and skill building over multiple sessions mediates the effects of OST setting participation on positive developmental outcomes.

Appendix D: Lower Stakes Accountability

In this appendix, higher and lower stakes accountability designs are compared. Figures D-1 and D-2 represent a "straw man" comparison of PBAS designs in state QRIS (Zellman et al., 2008) and city-wide QIS (Yohalem et al., 2012) respectively.

In brief, lower stakes approaches to accountability emphasize the experience of the individuals being held accountable. Higher and lower stakes refer to the experience of specific "accountabilities" or targeted performance levels linked with an explicit incentive. Lower stakes accountabilities are present when most individuals in the PBAS experience the specific accountabilities for their role as:

- Attainable with effort and support
- Worth the effort
- Precisely and fairly measured
- Focused on recognition for high performance.

Access to proven supports for attainment of performance levels decreases the stakes for a specific accountability because low performers are "rewarded" with additional supports and coaching. Higher-stakes accountabilities designs are characterized by greater emphasis on rewards for high performance. In the expanded learning field where the policy emphasis has been on building capacity of new providers, lower stakes models have helped to bring services to scale.

Figure D-1 describes a higher-stakes PBAS design that emphasizes a specific type of incentives, publicity of program. The PBAS theory described in Figure D-1 suggests: Performance on key service outputs is measured; ratings are made public so that consumers can choose high quality programs and undersubscribe low quality ones; Fear of low ratings will lead programs to take improvement action; overall service quality will improve.

Figure D-1: High-Stakes Accountability System Model for State QRIS[38]

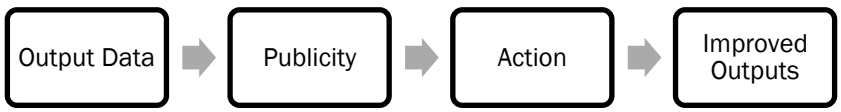
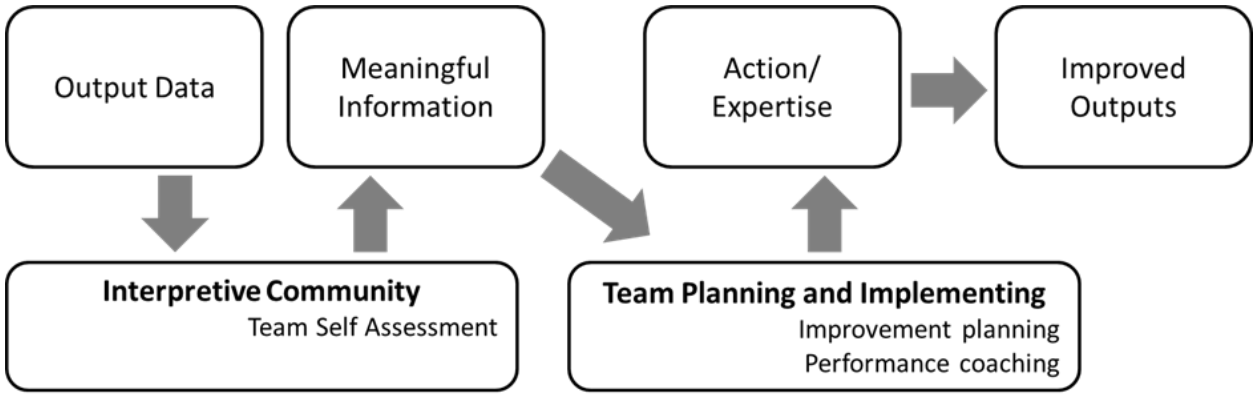


Figure D-2 describes a lower stakes PBAS design that emphasizes performance improvement supports. The PBAS theory described in Figure D-2 draws upon knowledge management theory (Mason,

[38] This logic model sequence was extracted from the left hand sequence of steps in Figure 1 (p.5) of Zellman et al. (2008). In fairness, their excellent discussion does reflect continuous improvement thinking (in the right hand sequence of Figure 1) but the QRIS's typically do not emphasize either lower stakes approaches or assure access to quality improvement supports which is the point of this contrast with PBAS in the expanded learning field.

2003) and research on data-driven improvement in schools (Halverson et al., 2007). The theory suggests that data have to be converted into meaningful information ("what it means to us") and then individuals need opportunities to learn and demonstrate expertise ("this is my plan") using information to change practice before improved outputs can be achieved. Notably, the sequence of knowledge management steps presented in Figure D-2 – data to information to expertise – are mediated by the sequence of performance improvement methods in the Youth Program Quality Intervention. Because many PBAS in the expanded learning field are based on the YPQI (or other similar interventions) as a site-level improvement model, we argue that the expanded learning field is a leading edge PBAS model.

Figure D-2: Lower-Stakes Accountability System Model

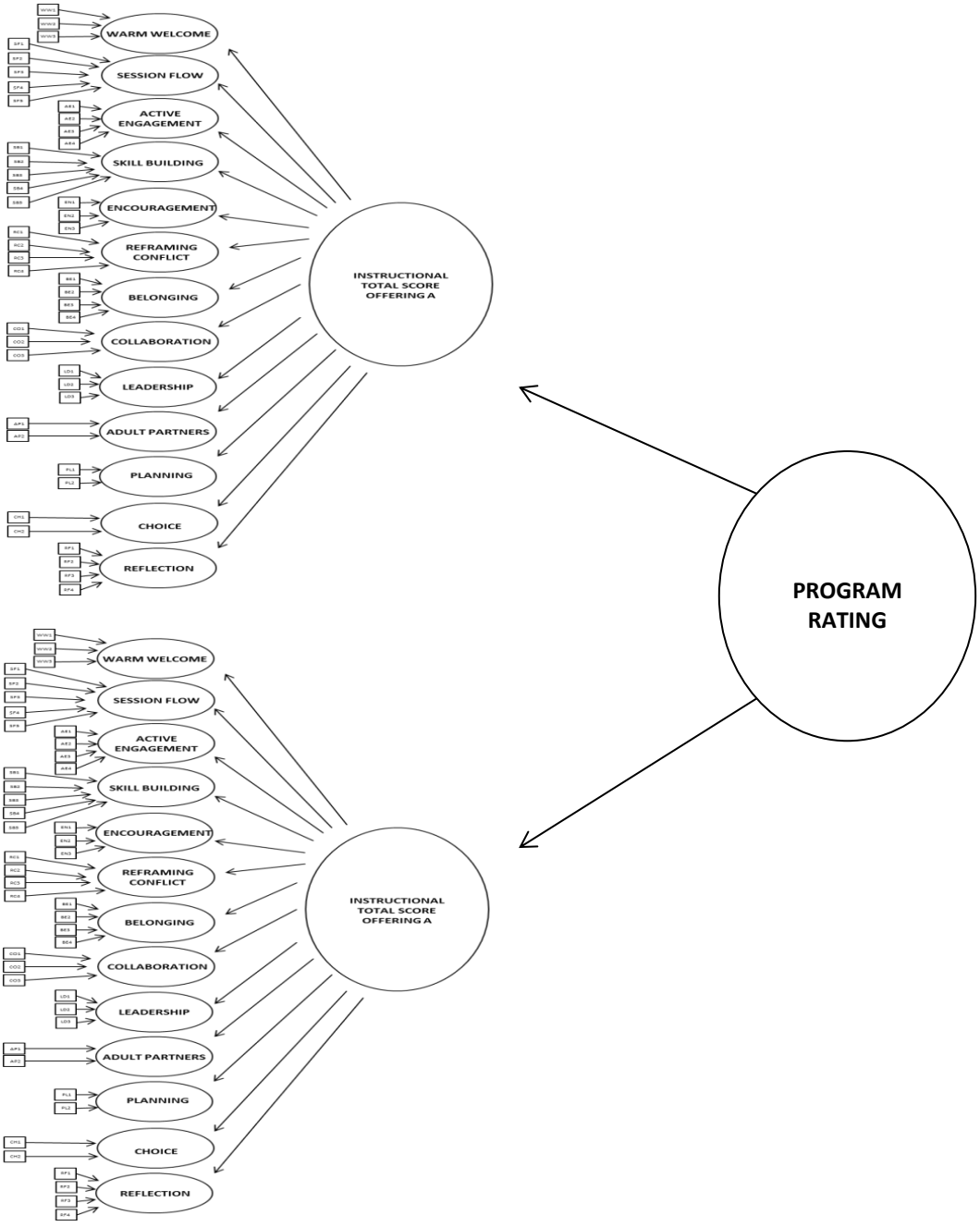


In the Quality Rating Improvement Systems (QRIS) policies for early childhood education, an explicit site-level improvement method is often lacking or under-emphasized. The National Child Care Information Center names the following characteristics of effective quality rating systems: (a) standards beyond licensing regulations, (b) accountability policies based on assessment and monitoring, (c) program and practitioner outreach and support, and (d) financing incentives specifically linked to compliance with quality standards (2009). Tout, Zaslow, Halle, and Forry (2009) suggested that the following factors may limit the impact of such policies: (a) small differences in structure and design (e.g., using different monitoring measures) make cross-site and network comparisons difficult; (b) coordination of improvement supports/momentum is blocked by lack of coordination across agencies, services, and data systems; and (c) policies lack clarity about goals, time frame, and expectations for actual improvement.

Appendix E: Program Rating Measurement Model

Figure E-1presents the measurement model for a program rating for instructional quality in an expanded learning PBAS that uses one of the Program Quality Assessment measures. Note that items are described as formative, while the instructional total score is described as a latent construct with reflective indicators, as is the program rating.

Figure E-1. Measurement Model for a Program Rating Based on Two Instructional Total Scores



Appendix F – Rasch Analyses Results[39]

In this technical appendix we utilize Rasch methods to examine the PBC-PQA data, specifically focusing on the issues of ceiling and floor effects and "naturally" occurring thresholds in skill levels for individual staff delivering instruction in Palm Beach expanded learning settings.

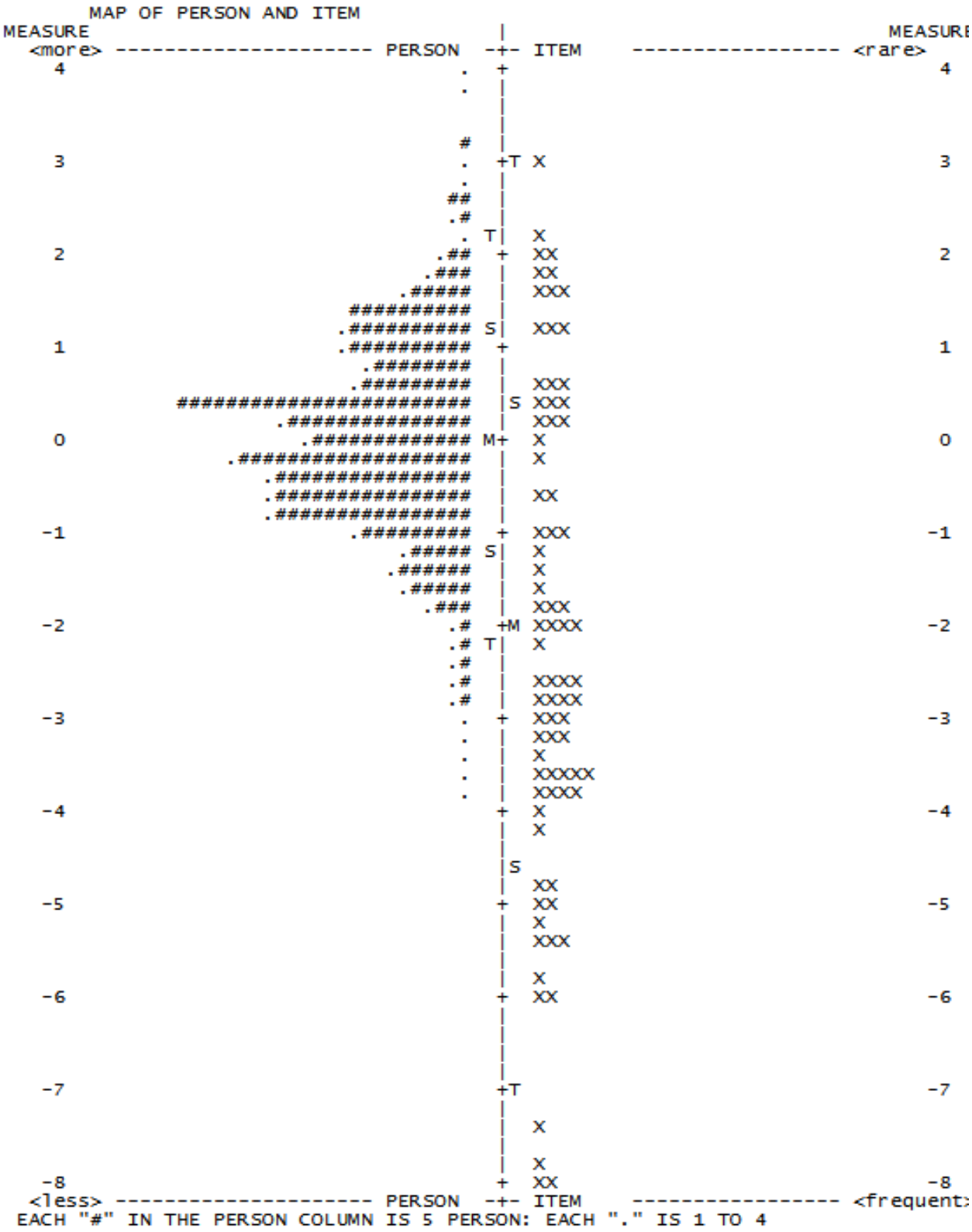
Ceiling and Floor Effects

In order to further assess the measurement properties of the PQA, steps were taken to calibrate PBC-PQA scale scores using Rasch analysis techniques. Application of Rasch modeling techniques yielded estimates of both a program's level of point-of-service quality and the relative difficulty of a given item appearing on the PBC-PQA. Working from the proposition that higher quality program would have had a greater likelihood of successfully obtaining a 5 on items represented on the PBC-PQA than lower quality programs, Rasch modeling techniques produced program quality and item difficulty estimates, transformed them using a log function, and allowed for each type of estimate to be displayed on a logit scale. This procedure allows program quality and item difficulty estimates to be directly compared on the same scale and one of outputs produced by the Winsteps software package is the variable map presented as Figure F-1. Variables maps can be employed to assess whether or not floor or ceiling effects are associated with a given scale and as a consequence, the degree to which the measure is likely to be sensitive to detecting offering session change over time.

In Figure F-1, all items contained on the PBC- PQA were employed to estimate a total score for each of the 1,428 program ratings in the Prime Time dataset (on the left side of the center line) and a difficulty estimate for each item (on the right side of the center line). Offerings on the lower end of the scale demonstrated a lower degree of instructional quality as defined by the PBC-PQA as compared to offerings near the top of the scale. Items at the low end of the scale were easier in the sense that raters had less difficulty rating them high (i.e., score of 5) than items near the top of the scale which were more difficult. Ideally, the mean of the offering session quality estimates and the mean of the item difficulties (each signified by M in the chart) would fall at a similar location along the scale. As outlined below, the mean of the item difficulty estimates is significantly lower than the mean of the offering session quality estimates. In this sense, the PQA is characterized by a lot of relatively easy items and the tool could benefit from some more difficult items or the elimination of some easier ones. However, there is no indication that floor or ceiling effects are in play here given that there is not a congregation of person estimates at either the top or bottom end of the scale.

[39] This appendix was prepared for the David P. Weikart Center by Neil Naftzger at American Institutes for Research.

Figure F-1. Variable Map for the PBC-PQA Total Score



Threshold Analysis

In this section we describe how Many Facet Rasch Measurement was used to further assess the psychometric properties of the PQA, as well as provide item difficulty estimates that were used to assess the viability of creating quality thresholds based on patterns of individual staff skills.

Many Facet Rasch Measurement

There are also Rasch analysis techniques, including Many Facet Rasch Measurement (MFRM), that can be employed to identify and quantify how other facets related to the measurement process may be impacting PQA scores. MFRM accomplishes this task by employing fit statistics and separation reliability indices to estimate parameters for a specific facet independent of the other facets included in the model. For example, the basic Rasch model allows for both the estimation of the quality of an offering and the difficulty of an individual item on the PQA and the production of individual standard errors for both quality and item difficulty estimates. MFRM allows a researcher to add additional facets to the Rasch model, like activity type, resulting in the estimation of how that facet is impacting quality estimates derived from the model. As noted by Kim and Wilson (2008), this feature of MFRM allows the researcher to assess the impact of error variance within each facet on the ability (or in this case, the quality) estimate. In this sense, the probability that an offering session will receive a given score on the measure of interest is a function of the difference between the quality of the offering session and the difficulty of the items, after adjusting for variation introduced by the type of activity observed. In this regard, as Kim and Wilson emphasize, what MFRM yields is an estimate of the quality of the offering session that is as free as possible from the particularities associated with the type of activity being observed.

Of the 1,428 offering level quality scores represented in the PrimeTime dataset, a total of 1,146 or 80 percent, had been coded as one of the following activity types:

- Homework Help/Tutoring
- Academic Enrichment
- Non-Academic Enrichment
- Recreation

It was hypothesized that enrichment-related activities would consistently score better on the PQA than homework help/tutoring or recreation activities. Running the MFRM calibrations demonstrated evidence to support this hypothesis. Activity type was shown to be significant as a factor, but the overall effect on

scores was relatively low. Nevertheless, application of this approach resulted in an adjustment to quality estimates to account for the type of activity being observed.

MFRM Estimates and Rating Scale Functioning

Rasch models also yield information about how well the rating scale associated with a given measure is functioning from a psychometric perspective. The MFRM procedures described in the previous section were used to assess PQA rating scale functioning from a Rasch perspective, yielding information about the actual width of each response option associated with a given rating scale relative to the construct being measured. Typically, ordinal response options akin to those found on the PQA are treated as covering an equal spectrum of the underlying construct of interest –in this case, quality practices supportive of youth development (that is, that the distance between a 1, 3, and 5 on the PQA are the same). However, when conducting Rasch analyses, the actual width of a response category is empirically based on how raters used the rating scale for the bank of items. The point where one rating option transitions to another is called a step calibration, or an Andrich threshold as it is referred to in the Figure F-2 below. As shown below, the transition point between a PQA rating of 1 and 2 (recoded from PQA rating of 3) occurs at -.12 logits while the transition point between 2 and 3 (recoded from PQA rating of 5) occurs at .12 logits. This indicated that the span of a PQA rating of 2 was found to be approximately .24 logits. According to Linacre (2004), the recommended minimum advance between step calibrations for a three category scale is 1.4 logits. As a consequence, based on the Prime Time dataset, raters used the PQA rating scale in a way that would suggest a two-point scale, where 1 and 3 are collapsed into one rating option and 5 is the other option, would produce a better functioning scale from a psychometric perspective. It may be advantageous in the future to redefine PQA rubrics associated with the middle score (a raw score of 3 recoded to a 2 for the analyses conducted here) to have wider substantive meaning. For the MFRM models described in this appendix, raw scores of 1 and 3 were collapsed into one category based on these findings.

Figure F-2. Andrich Thresholds for PQA Scores

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Response
Category Score	Counts Used	%	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	Thresholds Measure	S.E.	Measure at Category	-0.5	PROBABLE from	THURSTONE Thresholds	PEAK Prob	Category Name
1	9873	13%	13%	-.18	-.30	2.3			(-1.55)		low	low	100%	Low
2	12240	17%	30%	.43	.62	.5	-.12	.01	.00	-.86	-.12	-.53	36%	Medium
3	51696	70%	100%	2.80	2.78	1.0	.12	.01	(1.56)	.87	.12	.53	100%	High
									(Mean)		(Modal)	(Median)		

Using Item Difficulty Estimates to Support the Creation of Quality Thresholds

One of the benefits of Rasch-based approaches is that estimates of program quality derived from the PQA can be directly compared with item difficulty estimates. For example, item *II-I3- The activities provide all youth one or more opportunities to talk about (or otherwise communicate) what they are doing and what they are thinking about to others* was found to have an item difficulty estimate of .99 logits based on the MFRM-based calibrations previously described. If an offering session were found to have a quality score of .99 logits, then it would be possible to say that the probability of the offering session receiving a 5 as a score on item II-I3 would be 50 percent. If a different offering session received a quality score of say 1.2 logits, the probability that that offering session would receive a 5 would be greater than 50 percent since the quality (or ability estimate) exceeds the difficulty of the item. Conversely, if an offering session received a quality score of .6 logits, the probability of receiving a 5 on item II-I3 would be below 50 percent.

This characteristic of Rasch was used to order PQA items from easiest to hardest (as shown in the chart below) and then explore how different approaches to creating quality thresholds based on raw scores mapped against the probability that programs assigned to a given tier would demonstrate a greater than 50 percent probability of getting 5 on a given item.

As shown in the figure below, items appearing on the PBC-PQA have been ordered from easiest to hardest. It is important to note that there are 29 items appearing in the PQA that have been excluded from the chart below given that all programs represented in the dataset had very high probabilities of getting a 5 on these items. It is our sense that these items could possibly be considered for elimination from the PQA given the lack of variation in results.

Three approaches were used to create quality grouping based on PQA raw scores:

- Dividing program into quartiles based on their raw scores
- Identifying programs as either falling one standard deviation below the mean, as being within one standard deviation of the mean, or as being one standard deviation above the mean

- Employing quality cut scores derived from the practical experience of Prime Time quality advisors

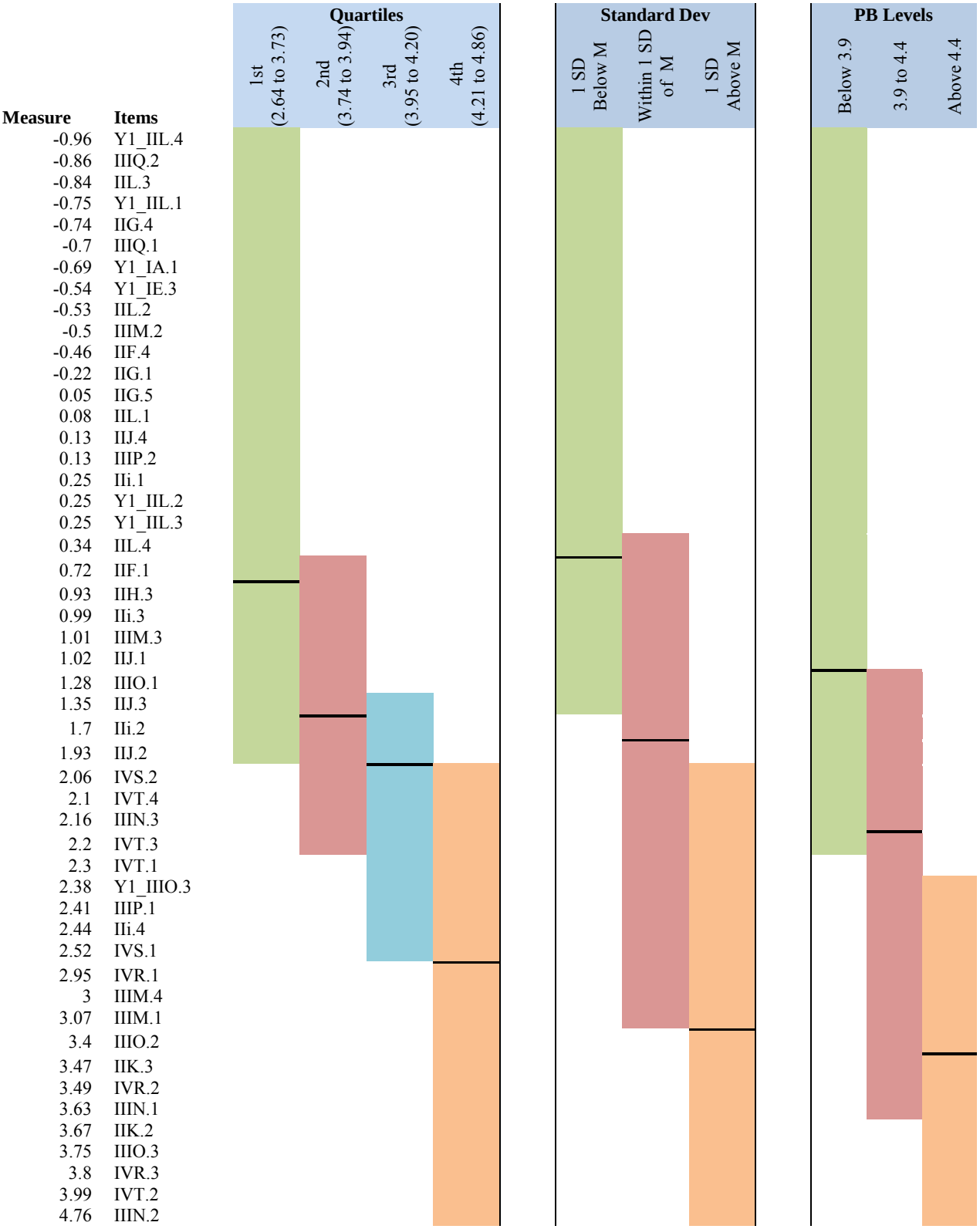
Using the mean raw total PQA score, quartiles were first created. Then, in each quartile column (see figure below), the range of logit scores associated with programs in that quartile were shaded and a solid blank line was drawn to show the approximate location of the logit mean. This quartile mean has substantive meaning. For example, if a program were found to have a quality estimate at the quartile 1 mean, it would have a greater than 50 percent probability of getting a 5 on all items between Y1_IIL.4 and IIF.1. Other quartile means could be interpreted in the same way.

This process was replicated for quality groups formed by employing the standard deviation and Prime Time-defined quality scores outlined in the figure below. The goal was to identify which raw score approach to creating quality categories seemed to yield substantively different grouping based on the practices that group was likely to have a greater than 50 percent probability of receiving a 5 on when the PQA was scored.

Overall, the standard deviation approach to creating thresholds yielded the most satisfying approach based on the methods employed. Threshold boundaries are defined by the mean logit values for programs assigned to a given group. As shown below, there appears to be a marked change in practices mastered in a given quality grouping. There is also a noticeable 'jump' in item difficulty from one grouping to the next. Additional efforts need to be undertaken to study the predictive value of thresholds formed in this manner relative to levels of youth engagement in programming and potentially other, more distal, program outcomes.

The following items were removed from these analyses: IA.2, IC.4, IC.5, IC.2, IA.1, IC.6, ID.4, IB.3, ID.2, IC.3, ID.3, ID.1, IE.1, IB.4, Y1_IA.2, Y1_IC.3, Y1_ID.4, IIG.3, IIG.2, IB.1, IE.3, Y1_ID.5, IE.2, IIH.1, IIK.1, IB.2, IIF.3, IIF.2, IC.1.

Figure F-3. PBC-PQA ordered from easiest to hardest



David P. Weikart
Center for Youth
Program Quality