AMERICAN INSTITUTES FOR RESEARCH®

Evaluation of Preschool for All (PFA) Implementation in San Francisco County

Year 5 Report

Submitted to:

First 5 San Francisco

Submitted by:

American Institutes for Research

September, 2010

2800 CAMPUS DRIVE, SUITE 200 | SAN MATEO, CA 94403 | TEL 650 843 8100 | FAX 650 858 0958 |
WEBSITE WWW.AIR.ORG

Acknowledgements

The American Institutes for Research (AIR) wants to extend thanks to the many individuals who contributed to this study, including the staff at First 5 San Francisco. We especially appreciate the preschool staff who participated in the classroom observations. The following AIR staff contributed to this report: Deborah Parrish, Principal Investigator; Gabriele Fain, Project Director; Jennifer Anthony, Research Analyst; Alison Hauser, Research Analyst; Shannon Madsen, Research Associate; and Connie Liu, Research Associate.

For more information about this report, please contact Gabriele Fain at:

American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Email: gfain@air.org
Phone: 650-843-8144

Executive Summary

The American Institutes for Research (AIR) conducted a five-year joint process evaluation, which began in December of 2005, to assess the implementation of Preschool for All (PFA), administered by First 5 San Francisco, in San Francisco County. The process evaluation was designed to investigate and document the implementation and the preliminary impacts of PFA on children, families, providers, and the community.

Each year of the study focused on emerging issues related to PFA implementation. The Year 5 evaluation consisted of two components. This first part of the study was designed to assess changes in the quality of PFA classrooms with teachers who participated in the *Institute for Intentional Teaching.* This professional development program was designed to strengthen teachers' instructional support strategies and was based on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). To measure changes in quality, AIR staff conducted pre- and post-CLASS observations of 11 PFA classrooms whose lead teachers were participating in the *Institute* program.

In the second component of the Year 5 evaluation, AIR staff pilot-tested a new research tool known as the Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA; Freedson, Figueras & Frede, 2008). Given the diversity of children served by PFA, First 5 San Francisco is particularly interested in tools that focus on the quality of instruction for dual language learners (DLLs). The CASEBA is designed to assess the degree to which preschool teachers and classrooms are providing support for the social, cognitive and linguistic development of dual language learners, with a focus on language and literacy. The CASEBA, like the CLASS, uses a 7-point Likert scale, with "1" representing the lowest quality, and "7" the highest.

The Year 5 pilot test yielded process information about the usability of the CASEBA, including its applicability in PFA's diverse classroom settings. To this end, classrooms in which the dominant language was Spanish and those in which Cantonese was the most common language were purposively sampled for observation. In addition, a pre- and post-test design was used for the pilot test, to explore whether the CASEBA was sensitive to change over time. Specifically, AIR staff observed nine classrooms, before and after the lead teacher participated in a professional development program related to supporting dual language learners.

Institute for Intentional Teaching

An overview of the results from the CLASS observations conducted in Year 5, which include dimension and domain scores, is shown in Exhibit A.

Exhibit A. Pre- and Post-Observation CLASS Scores

Domain	Dimension	2009 Pretest Average	2010 Posttest Average	Average Pre to Post Point Score Change+	Average 2009 Pretest Domain Scores	Average 2010 Posttest Domain Scores
Emotional Support	Positive Climate	6.25	6.43	.17	**Emotional Support** **6.2**	**Emotional Support** **6.3**
	Negative Climate[1]	1.16	1.12	-.04		
	Teacher Sensitivity	5.96	6.12	.16		
	Regard for Student Perspectives	5.55	5.63	.09		
Classroom Organization	Behavior Management	5.84	6.01	.18	**Classroom Organization** **5.5**	**Classroom Organization** **5.5**
	Productivity	5.75	5.81	.06		
	Instructional Learning Formats	4.88	4.60	-.27		
Instructional Support	Concept Development	1.86	2.13	.27	**Instructional Support** **2.6**	**Instructional Support** **3.3****
	Quality of Feedback	2.54	3.45**	.91		
	Language Modeling	3.40	4.38**	.98		

*p<.05, **p<.01, ***p<.001
+Rounding error

From the pre- to post-observations, the general pattern of CLASS remained the same, with high scores in the *Emotional Support* and *Classroom Organization* domains and lower scores for the *Instructional Support* domain. On average, sampled classrooms, on both of the pre- and post-observations, scored in the high-range on most of the CLASS dimensions. Classrooms demonstrated the highest quality in regard to their positive climate (and lack of negative climate), behavior management, and teacher sensitivity. The lowest ratings of quality were found for the dimensions focusing on concept development and quality of feedback to children. These ratings mirror those found in previous years of the PFA process evaluation.

All of the dimension scores increased in quality from the pre- to the post-observation (this includes *Negative Climate*, where a decrease in the score is actually a positive indication of quality improvement), with the exception of *Instructional Learning Formats*, which decreased by .27 points. The greatest increase was in the area of *Language Modeling* (a .98 point increase), followed by *Quality of Feedback* (.91 point increase). The increases in scores for both dimensions were statistically significant, as was the increase in the domain score that includes these dimensions, *Instructional Support*. *Concept Development*, the lowest scoring dimension in the pre-observations, remained the lowest among the dimensions in the post observations. This dimension increased by .27 points, from the pre- to post-test, yet remained in the "low" category of the CLASS.

The results indicate that PFA programs currently operating in San Francisco County typically offer warm and emotionally supportive teacher-child interactions. In addition, PFA teachers typically implement effective behavior and instructional management strategies to maximize

[1] The rating scale for Negative Climate is inversed – lower scores indicate higher quality.

learning opportunities for children. However, PFA teachers appear to be somewhat less effective in promoting children's higher-order thinking skills and cognition. However, considering pre and post-observations, there was statistically positive growth for two of the dimensions within the *Instructional Support* domain – almost a point increase for *Quality of Feedback* and *Language Modeling*, with both of these dimensions in the mid-range of the CLASS scoring system. These improvements are noteworthy because recent studies have found a relationship between program quality, as measured by the CLASS – and in particular *Instructional Support* – and children's academic outcomes. Howes et al. (2008) found that, among the CLASS domains, *Instructional Support* was the most consistent and robust dimension for predicting children's gains on receptive and expressive language assessments.

While this study was not designed to rigorously test the impact of the *Institute for Intentional Teaching* on teachers' use of effective instructional strategies, the findings do indicate that teachers who participated in the training program grew stronger in their use of effective instructional strategies over time.

Pilot Test of the CASEBA

Exhibit B shows the CASEBA pre- and post-scale results for the nine classrooms observed in the study. As noted earlier, each item on the CASEBA is rated on a 7-point Likert scale. A score of 7 indicates that a specific form of support and accompanying practices are present in a close-to-ideal form, and 1 represents the total absence of any such practices. A score of 1 on the scale indicates there is "no evidence" to support the item, a 3 represents "minimal evidence," a 5 is "good evidence," and a 7 is "strong evidence."

Exhibit B. Pre- and Post-Observation CASEBA Scale Scores

Scale (# of items in each scale)	Pre	Post	Difference
Supports for English Language Development (7 items)	4.81	4.62	-.19
Supports for Home Language Development (11 items)	3.90	3.79	-.11
Culturally Responsive Instruction (3 items)	5.63	5.41	-.22
Teacher Knowledge of Child Background (2 items)	5.17	6.11*	.94
Supports for Literacy Development in English (2 items)	4.22	4.22	0
Overall Score+ (26 items)	4.53	4.53	0

*p<.05

+The overall score is the mean of the 26 items, so each individual item is weighted the same, which is in line with NIEER's approach to calculating the overall CASEBA score. Each of the five scales is composed of a different number of individual items (e.g., *Culturally Responsive Instruction* is composed of three items, *Supports for Home Language Development* is made up of 11 items).

The CASEBA scale scores either decreased very slightly[2] or did not change from pre- to post-test, with the exception of *Teacher Knowledge of Child Background*, which increased by .94 points (a statistically significant increase). In general, CASEBA scores did not significantly increase for training participants after they completed the training. This may be a function of

[2] These decreases were not statistically significant.

several factors – the five month period between the pre- and post-observation may not have been long enough for teachers to adopt new practices, based on the training; the tool may not be sensitive to changes that do occur within such a time period, and/or the training did not align well with the CASEBA tool. The latter factor is most likely, as the training developers described the intervention as focusing on a few specific strategies, whereas the CASEBA is a more far-reaching tool related to program practices, classroom resources, as well as teacher behavior. In general, the CASEBA pilot test provided a snapshot of language and literacy practices to support DLL children at two points in time –and offered First 5 San Francisco some baseline information about how teachers and preschool settings support dual language learners.

In sum, the post-CASEBA observation scale scores indicated that:

- Programs and teachers demonstrated that they know the cultural and linguistic backgrounds of the children they serve (average score of 6.11).
- Classrooms also scored highly (average score of 5.41) in regard to incorporating the cultural backgrounds and life experiences of DLL children in the classroom, and in providing an emotionally warm and respectful environment for children.
- Classrooms scored in the mid range of the CASEBA rating scale (an average score of 4.62) in regard to teachers' use of high-quality talk in English and effective strategies to scaffold children's comprehension of instructional content in English.
- Classrooms scored similarly (average score of 4.22) in regard to the support of DLL's print literacy in English—this factor included indicators related to books, print, and literacy props in English and supporting children in learning print-related early literacy skills in English.
- The lowest score among the five scales was for *Supports for Home Language Development*, which received an average rating of 3.79. This factor includes a total of 11 items on the CASEBA, which relate to the use of home language for instructional purposes, the use of children's home language despite teachers' proficiency in the language, the use of high-quality talk in children's home language, and effective strategies to support children's development of their home language. In addition, the availability of books, print, and literacy props; teachers' support of the learning of print-related early literacy skills in the DLL children's home languages; and teachers' encouragement of DLL parents to maintain children's home language are included in this scale.

In general, AIR observers reported that the CASEBA was relatively easy to use and that they were able to observe and score items with confidence. The CASEBA is designed to be used in any preschool setting, regardless of the linguistic or cultural background of the children served. The tool was validated by NIEER in classrooms with a high proportion of Spanish-speaking DLL children. AIR staff generally felt the tool worked well in both types of classrooms they observed for this study (with Spanish-speaking and Cantonese-speaking DLLs). If First 5 San Francisco were to use the tool in the future, AIR recommends that they work in partnership with NIEER to refine the CASEBA training to include discussion of strong examples of best practices in PFA's Cantonese-dominant classrooms. Training serves several purposes—including identifying key examples observers may encounter in classrooms serving different cultural and linguistic groups and helping trainees to articulate their own assumptions and how they might influence their use of the tool. The small number of AIR observers for this study—three were used—enabled them to meet frequently to discuss their observations and seek clarification from NIEER on some items during the process.

In sum, the CASEBA provides rich detail on strategies and supports for DLL children in preschool settings, yet its format enables staff to administer the tool easily and within one observation session. In addition, the CASEBA can serve as a springboard for professional development efforts for teaching staff—most of the tool's items and indicators are easily understood and can inform training efforts.

Table of Contents

Year 5 PFA Process Evaluation

The American Institutes for Research (AIR) conducted a five-year joint process evaluation, which began in December of 2005, to assess the implementation of Preschool for All (PFA), administered by First 5 San Francisco, in San Francisco County. The process evaluation was designed to investigate and document the implementation and the preliminary impacts of PFA on children, families, providers, and the community.

Each year of the study focused on emerging issues related to PFA implementation. The Year 5 evaluation assessed the quality of PFA classrooms with teachers who participated in two professional development programs. They included the *Institute for Intentional Teaching,* designed to strengthen teachers' instructional support strategies and was based on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008). For this component of the study, AIR staff conducted pre- and post-CLASS observations of 11 PFA classrooms whose lead teachers were participating in the *Institute* program.

In the second component of the Year 5 evaluation, AIR staff pilot-tested a new research tool known as the Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA; Freedson, Figueras & Frede, 2008). Given the diversity of children served by PFA, First 5 San Francisco is particularly interested in tools that focus on the quality of instruction for dual language learners (DLLs). The CASEBA is designed to assess the degree to which preschool teachers and classrooms are providing support for the social, cognitive and linguistic development of dual language learners, with a focus on language and literacy. The Year 5 pilot test of the CASEBA yielded process information about the usability of the tool, including its applicability in PFA's diverse classroom settings. To this end, classrooms in which the dominant language was Spanish and those in which Cantonese was the most common language were purposively sampled for observation. In addition, a pre- and post-test design was used for the pilot test, to explore whether the CASEBA was sensitive to change over time. Specifically, AIR staff observed nine classrooms, before and after the lead teacher participated in a professional development program related to supporting dual language learners.

Institute for Intentional Teaching

First 5 San Francisco, in partnership with the San Francisco City College Early Childhood Mentor Program, hosted a professional development program for center directors, site supervisors, teachers, and assistant teachers at PFA sites. A cohort of PFA classrooms participated in a seven-month *PFA Institute for Intentional Teaching* during 2009-10. Specifically, the *Institute* focused on building quality programs based on the strategies assessed by the CLASS within the *Instructional Support* domain. The professional development program included: 1) training for program directors on their role as educational leaders, on the CLASS Pre-K framework for quality, and on how to support their teachers to enhance instructional support strategies, 2) monthly three-hour seminars on the CLASS instructional support strategies (see page 4 of this report for a full description of the dimensions included in the *Instructional Support* domain of the CLASS) for directors and teachers, 3) on-site technical assistance (one technical assistance visit per month, from October 2009 through April 2010) to support the teachers' implementation of the CLASS instructional support strategies, and 4) a practicum for teachers within the classroom involving an in-depth, child-centered project or study.

To assess whether the quality of teacher-child interactions improved after participating in the training program, AIR staff conducted pre- and post-observations using the CLASS. Eleven teachers – one from each preschool agency participating in the training – were randomly selected to be observed by AIR. The pre-observations were conducted in October of 2009. The post-observations were completed in the spring of 2010, between April and June. The data were analyzed to determine if there were significant changes in CLASS scores between the pre- and post-observations. Because resources did not allow for a comparison group in the study design, it is not possible to conclusively attribute changes in CLASS scores over time to teachers' participation in the professional development program. However, given the lack of improvement in CLASS scores across PFA classrooms in previous years of the PFA process evaluation, the results provide preliminary evidence to First 5 San Francisco regarding the effectiveness of the training program. In addition, the CLASS scores provide First 5 San Francisco with a continuing snapshot of the quality of PFA classrooms over time, with the Year 5 results building on observations conducted in Years 2 and 3 of the five-year study.

Overview of Classroom Assessment Scoring System

The CLASS has been used in Years 2, 3, and 5 of the PFA process evaluation to measure program quality. The tool builds on a broad body of research that highlights the critical nature of adult-child interactions in supporting children's learning and development. The CLASS framework measures adult-child interactions across several domains, including emotional and instructional support and classroom organization, drawing from the varied research base on teacher-child relationships, children's language and cognitive development, emotional and social functioning, self-regulatory skills, and classroom management practices. For example, researchers have found that teacher-child relationships are positively related to children's language skills and reading competence (Mashburn, Pianta, Hamre, Downer, Barbarin, Bryant, Burchinal, & Early, 2008; Burchinal, Peisner-Feinburg, Pianta, & Howes, 2002) and children's social competence (Mitchell-Copeland, 1997). The Cost, Quality & Outcomes Study (1999) indicated that children's cognitive development was positively related to the quality of classroom practices and that close teacher-child relationships were associated with better behavior and social skills through early elementary school. Hamre and Pianta (2005) found that students at risk of school failure who were enrolled in classrooms characterized by strong instructional and emotional support had higher achievement scores and lower levels of child-teacher conflict compared to children in less supportive environments. Underpinning the entire CLASS tool is the theory that the "primary mechanisms through which children acquire readiness-related competences are social relationships children form with peers, parents, and teachers" (Mashburn & Pianta, 2006).

The CLASS addresses three domains, *Emotional Support*, *Classroom Organization*, and *Instructional Support*, each consisting of one or more dimensions, as shown in Exhibit 1. Scoring on each dimension is based on observation of a series of indicators, also listed in Exhibit 1. Scoring for the CLASS dimensions is **not** determined by the presence of materials, the classroom's physical environment, safety issues, or a specific curriculum. Rather, the CLASS focuses on what teachers do with the materials they have and on staff-child interactions.

Exhibit 1. CLASS Domains, Dimensions, and Indicators

Emotional Support	
Dimensions	**Indicators**
Positive Climate. Reflects the emotional connection between the teacher and students and among students and the warmth, respect, and enjoyment communicated by verbal and nonverbal interactions.	• Relationships • Positive Affect • Positive Communication • Respect
Negative Climate. Reflects the overall level of expressed negativity in the classroom; the frequency, quality, and intensity of teacher and peer negativity are key to this scale.	• Negative Affect • Punitive Control • Sarcasm/Disrespect • Severe Negativity
Teacher Sensitivity. Encompasses the teacher's awareness of and responsivity to students' academic and emotional needs; high levels of sensitivity facilitate students' ability to actively explore and learn because the teacher consistently provides comfort, reassurance and encouragement.	• Awareness • Responsiveness • Addresses Problems • Student Comfort
Regard for Student Perspectives. Captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests, motivations, and points of view and encourage student responsibility and autonomy.	• Flexibility and Student Focus • Support for Autonomy and Leadership • Student Expression • Restriction of Movement
Classroom Organization	
Dimensions	**Indicators**
Behavior Management. Encompasses the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior.	• Clear Behavioral Expectations • Proactive • Redirection of Misbehavior • Student Behavior
Productivity. Considers how well the teacher manages instructional time and routines and provides activities for students so that they have the opportunity to be involved in learning activities.	• Maximizing Learning Time • Routines • Transitions • Preparation
Instructional Learning Formats. Focuses on the ways in which the teacher maximizes students' interest, engagement, and ability to learn from lessons and activities.	• Effective Facilitation • Variety of Modalities and Materials • Student Interest • Clarity of Learning Objectives
Instructional Support	
Dimensions	**Indicators**
Concept Development. Measures the teacher's use of instructional discussions and activities to promote students' higher-order thinking skills and cognition and the teacher's focus on understanding rather than on rote instruction.	• Analysis and Reasoning • Creating • Integration • Connections to the Real World
Quality of Feedback. Assesses the degree to which the teacher provides feedback that expands learning and understanding and encourages continued participation.	• Scaffolding • Feedback Loops • Prompting Thought Processes • Providing Information • Encouragement and Affirmation
Language Modeling. Captures the quality and amount of the teacher's use of language-stimulation and language-facilitation techniques.	• Frequent Conversation • Open-Ended Questions • Repetition and Extension • Self and Parallel Talk • Advanced Language

*Source: CLASS Manual, Preschool Version (2008)

Scoring the CLASS

The CLASS requires the observer to select a score for each of the 10 dimensions listed in Exhibit 1, based on the degree to which behavioral, emotional, and physical markers are observed and on the extent to which each dimension characterizes the classroom, rated from 1 (minimally characteristic) to 7 (highly characteristic). CLASS observations consist of three or more observation cycles. Each cycle includes a 20-minute observation period and a 10-minute period to record codes. To select a rating for each dimension, the observer must make judgments based upon the ranges of frequency, intention, and tone of interpersonal and individual behavior during the observation time.

The CLASS observations ran the entire length of the PFA session (approximately 3 to 3.5 hours), with the exception of outdoor play time, during which observations were not conducted.[3] At least four observation cycles (20-minute observations and 10-minute recording periods) were conducted at each program. When multiple teachers were in a classroom, teacher behaviors were weighted according to the number of children they worked with, the amount of time spent with children, and their responsibility for activities. Similar to the ECERS-R, the CLASS is meant to reflect the typical experiences for a child in the classroom.

CLASS scoring is completed immediately after each observation cycle. Each dimension is rated using a seven-point scale. Dimension descriptions at the “low,” “mid,” and “high” range are included in the CLASS manual and are used to select a rating. For example, after a 20-minute observation period that is guided by the indicators for each dimension (as shown in Exhibit 1) and in which notes are taken, the observer reads through the “low,” “mid,” and “high” range classroom descriptions that are included in the CLASS manual for each dimension. Once the appropriate level is selected, the observer then rates the dimension with a specific score (for “low” classrooms, a 1 or a 2; for “mid,” a 3, 4, or 5, etc.). The following rating scale provides guidance to observers in selecting the appropriate score for each dimension.

Exhibit 2. CLASS Rating Scale

Low		Mid			High	
1	**2**	**3**	**4**	**5**	**6**	**7**
The low range description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the low range are present.	**The low range description mostly fits the classroom/ teacher but there are one or two indicators that are in the mid range.**	**The mid range description mostly fits the classroom/ teacher but there are one or two indicators in the low range.**	**The mid range description fits the classroom/ teacher very well. All, or almost all, relevant indicators in the mid range are present.**	**The mid range description mostly fits the classroom/ teacher but there are one or two indicators in the high range.**	**The high range description mostly fits the classroom/ teacher but there are one or two indicators in the mid range.**	**The high range description fits the classroom/ teacher very well.**

*Source: CLASS Manual, Preschool Version (2008)

Pre- and Post-Observation CLASS Scores

An overview of the results from the CLASS observations conducted in Year 5, which include dimension and domain scores, is shown in Exhibit 3.

[3] The developers of the CLASS do not recommend conducting observations during outdoor play times, because it can be difficult to hear and observe staff-child interactions while teachers move around the outdoor space.

Exhibit 3. Average Pre and Post PFA CLASS Scores

Domain	Dimension	2009 Pretest Average	2010 Posttest Average	Average Pre to Post Point Score Change+	Range of Pre to Post Point Score Change	Average 2009 Pretest Domain Scores	Average 2010 Posttest Domain Scores
Emotional Support	Positive Climate	6.25	6.43	.17	-.75 – 1.00	**Emotional Support** **6.2**	**Emotional Support** **6.3**
	Negative Climate[4]	1.16	1.12	-.04	-1.00 – .40		
	Teacher Sensitivity	5.96	6.12	.16	-.83 – 1.50		
	Regard for Student Perspectives	5.55	5.63	.09	-.93 – 1.45		
Classroom Organization	Behavior Management	5.84	6.01	.18	-.60 – 1.25	**Classroom Organization** **5.5**	**Classroom Organization** **5.5**
	Productivity	5.75	5.81	.06	-.92 – 1.75		
	Instructional Learning Formats	4.88	4.60	-.27	-1.73 – 1.00		
Instructional Support	Concept Development	1.86	2.13	.27	-.25 – 1.40	**Instructional Support** **2.6**	**Instructional Support** **3.3****
	Quality of Feedback	2.54	3.45**	.91	-.50 – 2.60		
	Language Modeling	3.40	4.38**	.98	0 – 2.60		

*p<.05, **p<.01, ***p<.001
+Rounding error

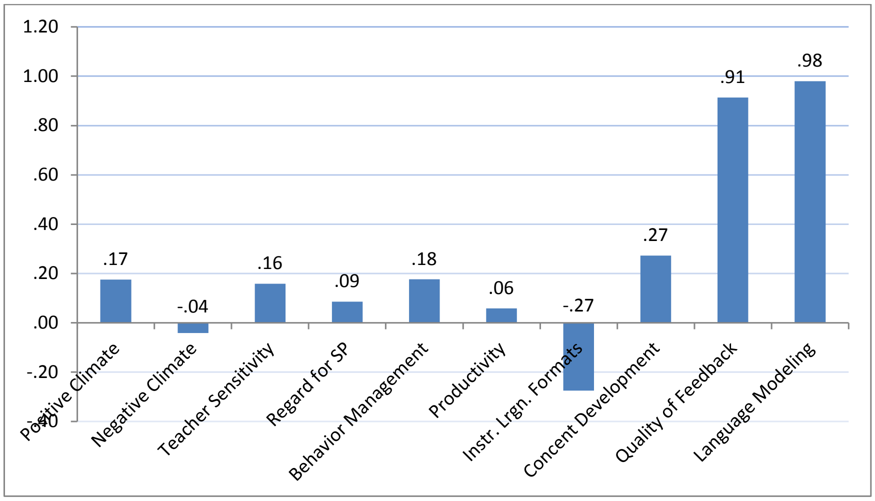
From the pre- to post-observations, the general pattern of CLASS remained the same, with high scores in the *Emotional Support* and *Classroom Organization* domains and lower scores for the *Instructional Support* domain. On average, sampled classrooms, on both of the pre- and post-observations, scored in the high-range on most of the CLASS dimensions. Classrooms demonstrated the highest quality in regard to their positive climate (and lack of negative climate), behavior management, and teacher sensitivity. The lowest ratings of quality were found for the dimensions focusing on concept development and quality of feedback to children. These ratings mirror those found in previous years of the PFA process evaluation.

All of the dimension scores increased in quality from the pre- to the post-observation (this includes *Negative Climate*, where a decrease in the score is actually a positive indication of quality improvement), with the exception of *Instructional Learning Formats*, which decreased by .27 points. The greatest increase was in the area of *Language Modeling* (a .98 point increase), followed by *Quality of Feedback* (.91 point increase). The increases in scores for both dimensions were statistically significant, as was the increase in the domain score that includes these dimensions, *Instructional Support*. *Concept Development*, the lowest scoring dimension in the pre-observations, remained the lowest among the dimensions in the post observations. This dimension increased by .27 points, from the pre- to post-test, yet remained in the "low" category of the CLASS.

[4] The rating scale for Negative Climate is inversed – lower scores indicate higher quality.

The average pre-to-post point differences for each of the 10 dimensions are shown in Exhibit 4.

Exhibit 4. Average Dimension Score Differences from Pre- to Post-Observation



Examining individual scores among the 11 classrooms, three classrooms (representing three different programs) demonstrated the largest gains, from pre- to post-observation. Classroom A showed the largest increases in three of the four dimensions within the *Emotional Support* domain (increases ranging from 1.00 to 1.5 points). For this classroom, scores were already in the “high” range for the pre-observation. Classroom B demonstrated the greatest increases pre to post in two of the three dimensions within the *Classroom Organization* domain – a 1.25 point increase for *Behavior Management* (although staying in the mid-range from pre to post) and a 1.75 point increase for *Productivity*, moving from the mid-range to the high-range on the CLASS rating system. Finally, Classroom C showed the greatest increases from pre- to post in the *Instructional Support* domain, with the highest average increases in two of the three dimensions – a 2.60 point increase in both *Quality of Feedback* and *Language Modeling*, moving from the low to mid-range in both of these dimensions.

Exhibit 5 shows the percentage of the 11 classrooms in the low, mid, and high range of the CLASS scoring system, on the pre- and post-observations. Average classroom scores for each dimension were rounded to the nearest whole number; scores of 1–2.4 fall into the low range, scores of 2.5–5.4 fall into the mid range, and scores of 5.5–7 fall into the high range.

Exhibit 5. Percentage of Pre and Post PFA Classrooms, by Score Category

	Pre-test 2009			Post-test 2010		
Dimension/Domain	**Low**	**Mid**	**High**	**Low**	**Mid**	**High**
Positive Climate	--	--	100%	--	--	100%
Negative Climate[5]	--	--	100%	--	--	100%
Teacher Sensitivity	--	27.3%	72.7%	--	--	100%
Regard for Student Perspectives	--	45.5%	54.5%	--	27.3%	72.7%
Emotional Support	--	9.1%	90.9%	--	--	100%
Behavior Management	--	27.3%	72.7%	--	18.2%	81.8%
Productivity	--	18.2%	81.8%	--	27.3%	72.7%
Instructional Learning Formats	--	81.8%	18.2%	--	72.7%	27.3%
Classroom Organization	--	63.6%	36.4%	--	45.5%	54.5%
Concept Development	90.9%	9.1%	--	81.8%	18.2%	--
Quality of Feedback	36.4%	63.6%	--	9.1%	90.9%	--
Language Modeling	27.3%	63.6%	9.1%	--	81.8%	18.2%
Instructional Support	45.5%	54.5%	--	18.2%	81.8%	--

Summary of Domain and Dimension Changes Over Time

The descriptions of low, mid, and high-range classrooms for each dimension are excerpted verbatim, with the author's permission, from the CLASS Preschool Manual (Pianta, La Paro, and Hamre, 2008). Given the nature of the CLASS scoring continuum, verbatim descriptors from the CLASS manual are used to ensure that the explanations for the San Francisco ratings accurately reflect the intent of the CLASS tool.

Emotional Support

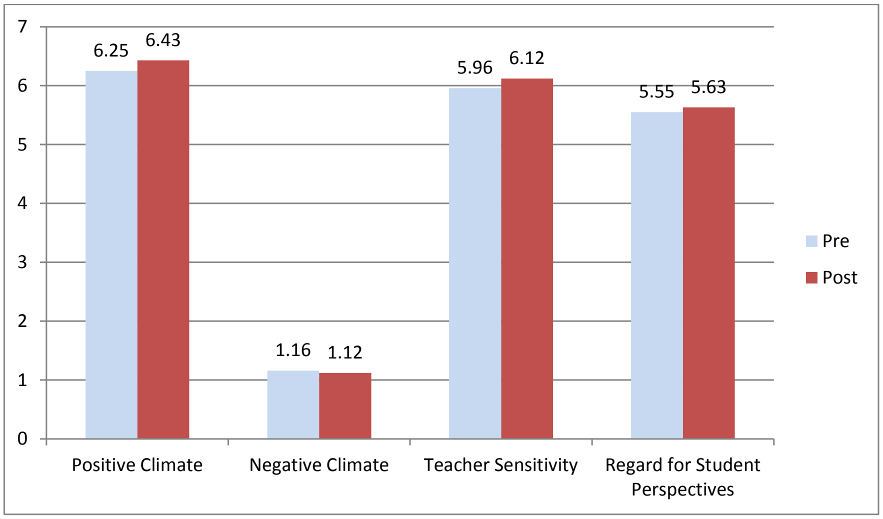
This domain reflects the emotional tone of the classroom, and includes measures of the positive and negative climate of the classroom, the extent to which teachers are sensitive to children, and teachers' regard for children's perspectives (e.g., focus on child autonomy and child-initiated activities). On the pre- and the post-test observations, the selected PFA classrooms received high scores – the overall domain score increased very slightly (from 6.2 to 6.3). All of the differences between the pre- and the post-test were positive (the exception being *Negative Climate*, where

[5] The rating scale for Negative Climate is inversed – lower scores indicate higher quality.

lower scores actually indicate higher quality), although none of the increases were statistically significant.

Exhibit 6 shows the average pre- and post-scores for the four dimensions that comprise the *Emotional Support* domain among PFA classrooms.

Exhibit 6. Pre and Post Average Scores for the Emotional Support Dimensions



All of the observed classrooms, both pre- and post-, scored in the high range for *Positive* and *Negative Climate*. Similarly, on the pre-test, most classrooms scored in the high range for *Teacher Sensitivity* (73% or eight classrooms), with the remainder (3 classrooms) in the mid-range. In the post-test, all classrooms scored in the high range for *Teacher Sensitivity*. In the typical high-range classroom, there were frequent displays of positive affect by the teacher and/or students. The classroom felt like a warm, pleasant place to be, with many instances of enthusiasm, including laugher or smiling among the teacher and students. There were frequently positive communications, verbal or physical, among teachers and students, and the teacher and students consistently demonstrate respect for one another.

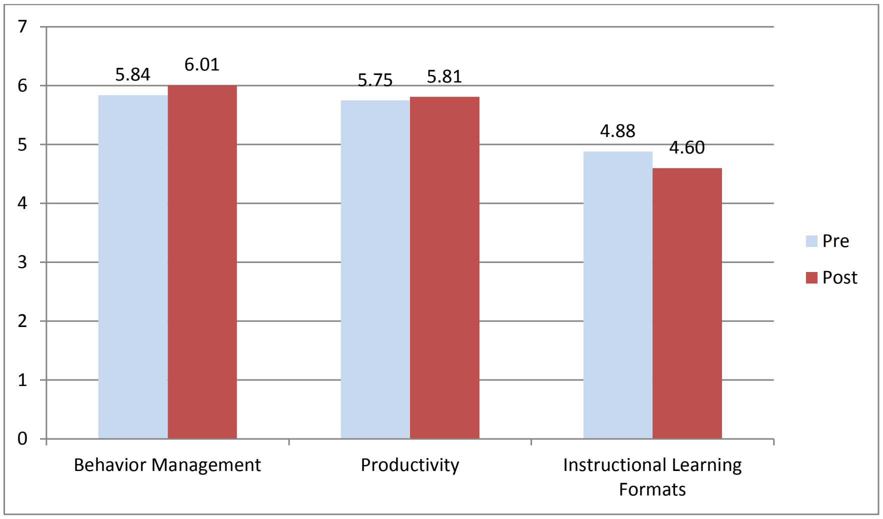
A similar pattern was found for *Regard for Student Perspectives*, with about half of the classrooms (5 classrooms) scoring in the mid-range on the pre-test and half (6 classrooms) in the high-range, with some increase on the post-test. There was some improvement on the post-test, with only three classrooms scoring in the mid-range and the rest (8 classrooms) in the high range. This dimension captures the degree to which the teacher's interactions with students and classroom activities place an emphasis on students' interests, motivations, and points of view and encourage student responsibility and autonomy. The typical mid-range *Regard for Student Perspectives* classroom is characterized by a teacher who may follow the students' lead during some periods and be more controlling during others. The teacher sometimes provides support for student autonomy and leadership but at other times fails to do so. There are periods during which there is a lot of student talk and expression but other times when teacher talk predominates. In addition, the teacher is somewhat controlling of students' movement and placement during activities. The pre- and post-observations showed some improvement across the classrooms, with more classrooms scoring in the high range for this dimension. In the typical high-range classroom for *Regard for Student Perspectives*, the teacher is flexible in his or her plans, goes along with students' ideas, and organizes instruction around students' interests. The teacher provides consistent support for student autonomy and leadership. There are many opportunities for student talk and expression. The students have freedom of movement and placement during activities.

In regard to *Teacher Sensitivity*, on the pre-test, 27% (three classrooms) scored in the mid range and 73% (eight classrooms) in the high range. On the post-test, 100% of classrooms scored in the high range. The typical classroom in the high range for this dimension has a teacher who is consistently aware of students who need extra support, assistance, or attention. The teacher is consistently responsive to students and matches his or her support to students' needs and abilities. The teacher is consistently effective at addressing students' problems and concerns. Finally, students appear comfortable seeking support from, sharing their ideas with, and responding freely to the teacher. In the typical mid-range *Teacher Sensitivity* classroom – such as the three classrooms that scored in this category on the pre-test – these strategies are not implemented consistently. For example, a teacher may seem very clued in to students' academic needs, giving them appropriate tasks, supporting their learning, and so forth, but less aware of students' emotional functioning. Or, a teacher may show elements of responsiveness, but at times miss or ignore students' attempts to get his or her attention.

Classroom Organization

The *Classroom Organization* domain reflects the effectiveness of the teacher's behavior management strategies, the extent to which children have opportunities to be involved in learning activities, and what the teacher does to maximize children's interest, engagement, and ability to learn from lessons and activities. Exhibit 7 shows the average pre- and post-scores for the three dimensions that comprise the *Classroom Organization* domain: *Behavior Management, Productivity*, and *Instructional Learning Formats*.

Exhibit 7. Pre and Post Average Scores for the Classroom Organization Dimensions



On the pre-test, most of the sampled PFA classrooms fell into the high range for *Behavior Management* and *Productivity* (73% and 82% of classrooms, respectively). According to the high-range CLASS descriptors for *Behavior Management*, rules and expectations for behavior are clearly and consistently enforced. The teacher is consistently proactive and monitors the classroom effectively to prevent problems from developing (e.g., teachers always appear to be one step ahead of problems in the classroom, anticipating and preventing misbehavior). The teacher effectively redirects misbehavior by focusing on positives and making use of subtle cues. Behavior management does not take time away from learning. In addition, there are few, if any, instances of student misbehavior. On the post-test for this dimension, scores tended to stay approximately the same – for *Behavior Management*, one classroom moved from the mid to high range.

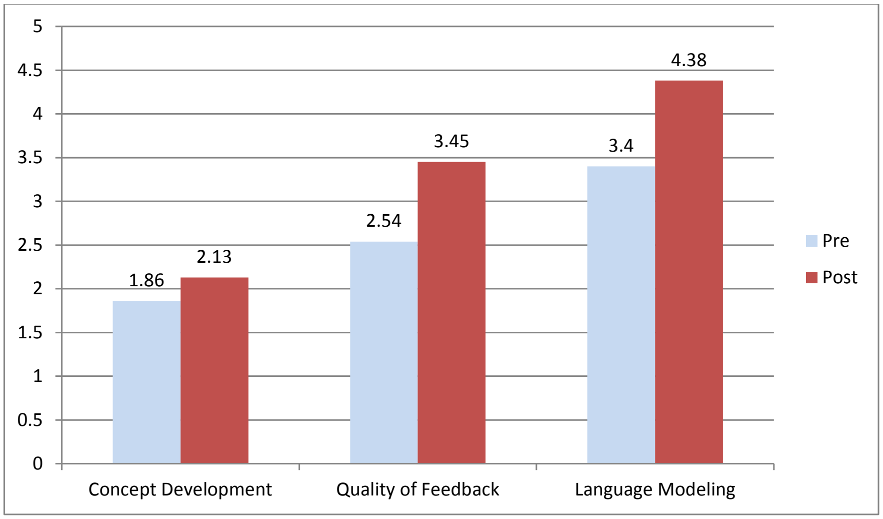
In regard to the typical high-range *Productivity* classroom, the teacher provides activities for the students and deals efficiently with disruptions and managerial tasks. The classroom is a “well-oiled machine”; everybody knows what is expected of them and how to go about doing it. Transitions are quick and efficient and the teachers are fully prepared for activities and lessons. On the pre-test, 82% or nine classrooms scored in the high range – on the post-test, one of these classrooms moved down to the mid range.

The final dimension included in the *Classroom Organization* domain is *Instructional Learning Formats*, for which the majority of classrooms on both the pre and post test scored in the mid range (82% on the pre-test and 73% on the post test). Based on the CLASS descriptors, the teacher in a typical mid-range classroom for *Instructional Learning Formats* actively facilitates activities and lessons to encourage interest and expanded involvement but at other times merely provides activities for the students. The teacher is inconsistent in his or her use of a variety of modalities and materials to gain students' interest and participation during activities and lessons. Students may be engaged and/or interested for periods of time, but at other times their interest wanes and they are not involved in the activity or lesson. In addition, the teacher orients students somewhat to learning objectives, or the learning objectives may be clear during some periods, but less so during others. In the high-range classroom (18% of classrooms on the pre-test and 27% on the post-test), the teacher actively facilitates students' engagement in activities and lessons to encourage participation and expanded involvement. The teacher uses a variety of modalities including auditory, visual, and movement and uses a variety of materials to effectively interest students and gain their participation during activities and lessons. Students are consistently interested and involved in activities and lessons. The teacher effectively focuses students' attention toward learning objectives and/or the purpose of the lesson.

Instructional Support

The lowest average domain score across PFA classrooms was for *Instructional Support*, on both the pre-test and the post-test. However, there was a statistically significant increase in the domain score across the two observations, from 2.6 to 3.3. *Instructional Support* reflects the teachers' use of instructional discussions and activities to promote children's higher-order thinking skills and cognition and the teachers' focus on understanding rather than on rote instruction, the degree to which the teachers provide feedback that expands learning and understanding and encourages continued participation, and the quality and amount of teachers' use of language-stimulation and language-facilitation techniques. The Instructional Support domain consists of three dimensions: *Language Modeling*, *Quality of Feedback*, and *Concept Development*.

Exhibit 8. Pre and Post Average Scores for the Instructional Support Dimensions



Across all the dimensions on the CLASS, the greatest increase was observed for two dimensions within the *Instructional Support* domain – *Language Modeling* and *Quality of Feedback*. Both increases were statistically significant. On the pre-test for *Language Modeling,* classrooms were spread out across the low (3 classrooms), mid (7 classrooms), and high ranges (one classroom) of the CLASS scoring system. On the post-test, none of the classrooms scored in the low range, with the majority (nine classrooms) in the mid-range and two in the high range. In the typical mid-range *Language Modeling* classroom, the teacher talks regularly with and to students and appears somewhat interested in students; however, conversations typically are limited to one or two back-and-forth exchanges rather than developing into prolonged conversations. The teacher asks a mix of closed-ended and open-ended questions. The teacher sometimes asks questions that require the students to use more complex language; however, the majority of questions are close-ended and require only short responses from the students. He or she sometimes repeats or extends students' responses and occasionally maps his or her own actions and the students' actions through language and description. Finally, the teacher sometimes uses advanced language with students. In the high-range classroom, these strategies are used more consistently by teachers.

In regard to *Quality of Feedback*, a somewhat similar pattern from pre- to post-test was found. On the pre-test, four classrooms scored in the low range, and the remainder (7 classes) scored in the mid-range of the CLASS. PFA classrooms that receive a score in the low range for *Quality of Feedback* typically are characterized by a teacher who rarely provides scaffolding to students but rather dismisses responses or actions as incorrect or ignores problems in understanding. The teacher in this classroom gives only perfunctory feedback to students, rarely queries the students or prompts them to explain their thinking and rationale for responses and actions, and rarely provides additional information to expand on the students' understanding or actions.

By the post test, all but one classroom scored in the mid-range for this dimension. In the typical mid-range classroom for *Quality of Feedback*, the teacher occasionally provides scaffolding to students but at other times simply dismisses responses as incorrect or ignores problems in students' understanding. There are occasional feedback loops – back-and-forth exchanges – between the teacher and students; however, feedback is more perfunctory. At times, the teacher's feedback may help students to expand and elaborate on their learning, but generally these efforts by the teacher are not sustained for long. More often, the teacher simply suggests that the students' answers are feasible (e.g., "That was a good guess. Does anyone else have an idea?") and then moves on to another student. The teacher occasionally queries the students or prompts students to explain their thinking and rationale for responses and actions. In response to student comments or actions, the teacher occasionally will ask *why* questions to prompt the student to explain his or her thinking and describe his or her actions; however, this does not occur often or is typically a very brief exchange. The teacher occasionally provides additional information to expand on the students' understanding or actions. For example, the teacher sometimes goes beyond perfunctory feedback (i.e., saying that a child's response is correct or incorrect), but this is not his or her typical style of response. Finally, the teacher occasionally offers encouragement of students' efforts that increases students' involvement and persistence.

The last dimension within the *Instructional Support* domain, *Concept Development*, showed a slight increase from pre- to post-test (.27), although the gain was not statistically significant. The average score for *Concept Development* remained the lowest across all dimensions, for both sets of observations. On the pre-test, 10 of the 11 classrooms scored in the low range, and one in the mid-range. On the post-test, 9 of the 11 classrooms continued to score in the low range, and two scored in the mid range, meaning that one classroom moved from the low to mid-range. In the typical low-range classroom for *Concept Development*, the teacher rarely uses discussions and activities that encourage analysis and reasoning. The teacher makes no attempt to develop students' understanding of ideas and concepts; the preponderance of teaching is focused on getting students to remember and repeat facts and practice basic skills. The teacher rarely provides opportunities for students to be creative and/or generate their own ideas and products. The focus in this classroom is on having students do things in a particular way rather than on helping to stimulate their creativity and ability to plan. The teacher fails to make use of brainstorming as a way to get students thinking. For example, after reading a book, the teacher asks questions such as "What did the frog need to build his house?" with the explicit goal of having students recall a fact from the story. (Instead, the teacher could have asked, "What *could* a frog use to build a house?" allowing students to think of many possible responses.) Concepts and activities are presented independent of one another, and students are not asked to apply previous learning. In this classroom, the teacher moves from one distinct subject to another and makes no attempt to link concepts. The teacher does not relate concepts to students' actual lives. Activities and instruction in this classroom seem abstract and removed from the students' everyday lives. The teacher does not provide opportunities for the students to apply knowledge to meaningful activities. For example, a teacher may conduct a lesson on the letter *t* but focus only on what a *t* looks like and how it sounds. He may make no attempts to have students look around the classroom for items that begin with *t*, generate a list of their own words that start with a *t*, or think of everyone in the room who has a t in their name. For those two classrooms that received a mid-range score for *Concept Development*, classrooms activities and discussions are used to a greater extent (although not consistently) to promote students' higher-order thinking skills and cognition.

Comparison CLASS Data

In Year 2 (2007) and Year 3 (2008) of the PFA process evaluation, AIR staff conducted CLASS observations in a representative sample (n=32 in 2007 and n=27 in 2008) of all PFA classrooms in San Francisco County. These results are shown in comparison with the 2009 and 2010 CLASS data (n=11), in Exhibit 8. However, it is important to note that the 2007 and 2008 sample was randomly selected from all PFA programs, whereas the 2009/10 sample was purposively drawn from programs that were participating in the *Institute for Intentional Teaching*. With the data presented in Exhibit 9, it is not possible to compare change over time within the same PFA classrooms. The data should not be interpreted as rigorous evidence regarding changes in the quality of all PFA classrooms in the county over time.

Exhibit 9. Average 2008, 2009, and 2010 San Francisco PFA CLASS Scores[6]

Domain	Dimension	2007 Average Scores	2008 Average Scores	2009 (Pretest) Average Scores	2010 (Posttest) Average Scores	2007 Average Domain Scores	2008 Average Domain Scores	2009 (Pretest) Average Domain Scores	2010 (Posttest) Average Domain Scores
Emotional Support	Positive Climate	6.19	5.93	6.25	6.43	**6.0**	**5.8**	**6.2**	**6.3**
	Negative Climate	1.24	1.22	1.16	1.12				
	Teacher Sensitivity	5.48	5.44	5.96	6.12				
	Regard for Student Perspectives	5.64	4.89	5.55	5.63				
Classroom Organization	Behavior Management	5.90	5.69	5.84	6.01	**5.2**	**5.3**	**5.5**	**5.5**
	Productivity	5.64	5.55	5.75	5.81				
	Instructional Learning Formats	3.90	4.76	4.88	4.60				
Instructional Support	Concept Development	2.88	1.95	1.86	2.13	**3.7**	**2.8**	**2.6**	**3.3**
	Quality of Feedback	3.40	2.48	2.54	3.45				
	Language Modeling	4.70	3.96	3.40	4.38				

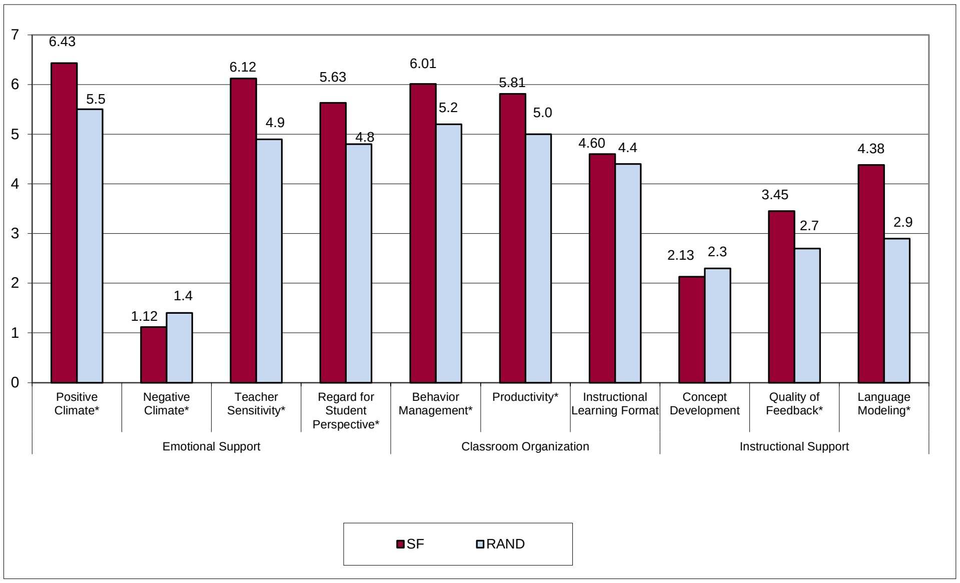
As shown in the table, scores for the sample in 2010 were somewhat higher than the sample of classrooms in 2008, with the greatest increases seen for the *Quality of Feedback* (.97 difference), *Regard for Student Perspectives* (.74 difference), and *Teacher Sensitivity* (.68). When compared

[6] It is important to note that the 2007 and 2008 sample was randomly selected from all PFA programs, whereas the 2009/10 sample was purposively drawn from programs that were participating in the *Institute for Intentional Teaching.*

to the 2007 scores, the 2010 post-scores are about the same – with a slightly higher average score for *Teacher Sensitivity* (.64) and slightly lower average score for *Concept Development* (.75). Again, it is important to note that unlike the 2007 and 2008 data, the 2009-2010 data do not reflect a representative sample of classrooms but rather reflect a purposively selected sample of classrooms whose teachers were participating in the *Institute for Intentional Teaching*. In regard to the 2008 dimension scores in the *Instructional Support* domain, it is not clear why they were lower than the 2007 scores – it could be due to the fact that the first programs to implement PFA were those most "ready" (i.e., most likely to meet PFA's quality criteria and subsequent years included more programs that needed more technical assistance to reach those standards).

In addition, PFA scores from the 2009/2010 sample are higher than statewide CLASS data, as was the case in previous years of the PFA process evaluation. The RAND Corporation conducted the California Preschool Study to examine the adequacy and efficiency of preschool education in the state. The 2008 study involved 615 observations among center-based programs in California – CLASS results are shown in Exhibit 10 and compared to the 2010 (post-test) San Francisco results.

Exhibit 10: RAND Preschool Study and 2010 San Francisco PFA CLASS Scores



*p<.05

The pattern of San Francisco PFA CLASS scores is very similar to that of the scores collected through the RAND study, with overall higher scores in the *Emotional Support* and the *Classroom Organization* domains and lower scores for the *Instructional Support* domain. The 11 PFA classrooms observed in 2010 showed the greatest differences, compared to the RAND data, in *Language Modeling* (a 1.48 point difference), followed by *Teacher Sensitivity* (1.22 point difference), and *Positive Climate* (.93 difference). All of the differences between San

Commented [GF1]: Phil – does this little footnote look right? should it be right under the table? It looks like the spacing is off between the table, this note, and the following paragraph, but I'm not sure how to fix.

Francisco and the RAND classroom dimension scores are statistically significant (p<.05), with the exception of *Instructional Learning Formats and Concept Development*.

Conclusion

The results indicate that PFA programs currently operating in San Francisco County typically offer warm and emotionally supportive teacher-child interactions. In addition, PFA teachers typically implement effective behavior and instructional management strategies to maximize learning opportunities for children. However, PFA teachers appear to be somewhat less effective in promoting children's higher-order thinking skills and cognition. Within the *Instructional Support* domain, the *Concept Development* dimension received the lowest score (2.13) on the post-test, a .27 increase from the average pre-test score of 1.86. Specifically, *Concept Development* focuses on analysis and reasoning (e.g., why and/or how questions, problem solving, prediction/experimentation, classification/comparison, evaluation); creating (e.g., brainstorming, planning, producing); integration (e.g., connects concepts, integrates with previous knowledge); and connections to the real world (e.g., real-world applications, related to students' lives). *Concept Development* has been the lowest-scoring dimension in observations conducted over the course of the PFA process evaluation – it seems to be persistently challenging for teachers to demonstrate mid- to higher-level scores on this dimension. This is true in San Francisco County PFA classrooms, as well as the state as a whole, as demonstrated by the RAND study.

Considering pre and post-observations, however, there was statistically positive growth for two of the dimensions within the *Instructional Support* domain – almost a point increase for *Quality of Feedback* and *Language Modeling*, with both of these dimensions in the mid-range of the CLASS scoring system. These improvements are noteworthy because recent studies have found a relationship between program quality, as measured by the CLASS – and in particular *Instructional Support* – and children's academic outcomes. Howes et al. (2008) found that, among the CLASS domains, *Instructional Support* was the most consistent and robust dimension for predicting children's gains on receptive and expressive language assessments.

As noted earlier, First 5 San Francisco resources did not allow for a comparison group in the study design, so it is not possible to conclusively attribute changes in CLASS scores over time to teachers' participation in the professional development program. However, the findings do indicate that teachers who participated in the *Institute for Intentional Teaching* grew stronger in their use of effective instructional strategies over time.

Pilot Test of the Classroom Assessment of Supports for Emergent Bilingual Acquisition

Preschool for All in San Francisco County serves a diverse group of children – in the 2008-2009 program year, 66 percent of PFA children had a home language other than English. Twenty-nine percent of all children spoke Cantonese at home, while 27 percent spoke Spanish at home. First 5 San Francisco is interested in research tools to assess the quality of supports provided to dual language learners in PFA classrooms. In Year 3 of the PFA Evaluation, AIR and First 5 San Francisco pilot tested a new tool known as the Language Interaction Snapshot (LISn), developed by Mathematica Policy Research, which provided information on the languages spoken in preschool settings and the types of verbal communications occurring between teachers and children.

In Year 5, another tool was pilot tested – the Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA; Freedson, Figueras, & Frede, 2008). The CASEBA is a newly developed research tool designed to assess the degree to which preschool teachers and classrooms are providing support for the social, cognitive and linguistic development of dual language learners (DLLs), with a focus on language and literacy. The instrument consists of 26 items, clustered around five scales: *Supports for English Language Development*, *Supports for Home Language Development*, *Culturally Responsive Instruction*, *Teacher Knowledge of Child Background*, and *Supports for Literacy Development in English.*[7] Each item is rated on a 7-point Likert scale, where 7 indicates that a specific form of support and accompanying practices are present in a close-to-ideal form, and 1 represents the total absence of any such practices. A score of 1 on the scale indicates there is “no evidence” to support the item, a 3 represents “minimal evidence,” a 5 is “good evidence,” and a 7 is “strong evidence”.

The Year 5 pilot test was designed with several goals in mind. The study yielded process information about the usability of the tool itself, including its applicability in PFA’s diverse classroom settings. To this end, classrooms in which the dominant language was Spanish and those in which Cantonese was the most common language were purposively sampled for observation. In addition, a pre- and post-test design was used for the pilot test, to explore whether the CASEBA was sensitive to change over a short time period (approximately five months).

Specifically, AIR staff observed nine classrooms, before and after the lead teacher participated in a professional development program related to supporting dual language learners. The training involved interactive seminars, on-site coaching, and mentoring of coaches to build the capacity of San Francisco early childhood professionals to support DLL children’s learning. The 2010 training was not limited to PFA teaching staff—it also included teachers employed by programs that were not currently being funded by PFA. It is important to note that the intent of the pilot test was not to evaluate the effectiveness of the professional development program. Rather, the teachers in the pilot study sample were known to have a high proportion of DLL children in their

[7] In fact, 25 of the 26 items map onto the five identified scales. Item #26, which relates to assessment of DLL children, does not map onto any of the five scales based on NIEER’s analysis.

classrooms and the pre/post study design also provided an opportunity to test the tool's sensitivity to change over time.

In six of the observed classrooms, the majority of children were Spanish-speaking DLL students, and in the other three classrooms, Cantonese was the most common home language. The pre-observations were conducted during a two-week window in January of 2010, before the training began, and the post-observations were completed between April and June of the same year, after the training was completed.

Pre- and Post-Observation CASEBA Scores

Exhibit 11 shows the CASEBA pre- and post-scale results for the nine classrooms.

Exhibit 11. Pre- and Post-Observation CASEBA Scale Scores

Scale (# of items in each scale)	Pre	Post	Difference
Supports for English Language Development (7 items)	4.81	4.62	-.19
Supports for Home Language Development (11 items)	3.90	3.79	-.11
Culturally Responsive Instruction (3 items)	5.63	5.41	-.22
Teacher Knowledge of Child Background (2 items)	5.17	6.11*	.94
Supports for Literacy Development in English (2 items)	4.22	4.22	0
Overall Score+ (26 items)	4.53	4.53	0

*p<.05

+The overall score is the mean of the 26 items, so each individual item is weighted the same, which is in line with NIEER's approach to calculating the overall CASEBA score. Each of the five scales is composed of a different number of individual items (e.g., *Culturally Responsive Instruction* is composed of three items, *Supports for Home Language Development* is made up of 11 items).

The scale scores either decreased very slightly[8] or did not change from pre- to post-test, with the exception of *Teacher Knowledge of Child Background*, which increased by .94 points (a statistically significant increase). This scale includes two CASEBA items (Item 1, *Systematic information on the language and cultural background of each child in the classroom is collected and available at the school/center site*, and Item 2, *The lead teacher knows the language and cultural background of each child in the classroom*).

[8] These decreases were not statistically significant.

Exhibit 12 shows the mean pre- and post-scores for the nine classrooms for each of the 26 CASEBA items.

Exhibit 12. Pre- and Post-Observation CASEBA Item Scores

Scale[9]	Item	Description	Pre	Post	Difference
Teacher Knowledge of Child Background	1	Systematic information on the language and cultural background of each child in the classroom is collected and available at the school/center site.	4.78	5.78	1.00
	2	The lead teacher knows the language and cultural background of each child in the classroom.	5.56	6.44	0.88
Culturally Responsive Instruction	3	The cultural backgrounds and life experiences of the DLL children are incorporated into the life of the classroom.	3.78	3.56	-0.22
Supports for Home Language Development	4	The lead teacher uses a home language of the DLL children for instructional purposes.	4.33	3.89	-0.44
	5	The paraprofessional or assistant teacher uses a home language of the DLL children for instructional purposes.	5.11	5.44	0.33
	6	The lead teacher attempts to learn and use the home language(s) spoken by the DLL children in the classroom, although she/he lacks proficiency in the language.	2.25	2.60	0.35
	7	The lead teacher uses high-quality talk in the children’s home language.	4.11	3.22*	-0.89
	8	The assistant teacher uses high-quality talk in the children’s home language.	5.00	4.22*	-0.78
	9	Teaching staff use effective strategies during group instruction to support DLL children’s development of their home language(s).	4.33	2.89**	-1.44
	10	Teaching staff interact one on one with DLL children in ways that support development of the home language.	4.78	3.78*	-1.00
	11	Teaching staff expand children’s repertoire of concepts and vocabulary in the home language.	3.11	2.78	-0.33
	12	Books, print, and literacy props are available in the DLL children’s home language/s.	3.56	4.22	0.66
	13	Teaching staff support the learning of print-related early literacy skills in the DLL children’s home language(s).	1.33	1.67	0.34
	14	Teaching staff encourage DLL parents to maintain children’s home language.	3.78	6.22**	2.44
Supports for English Language Development	15	The lead teacher uses high-quality talk in English.	5.33	5.11	-0.22
	16	The assistant teacher uses high-quality talk in English.	4.00	3.78	-0.22
	17	Teaching staff use effective strategies to scaffold children’s comprehension of instructional content in English.	6.22	5.89	-0.33
	18	Teaching staff use effective strategies during group instruction to build children’s communication skills in English.	5.00	3.89*	-1.11

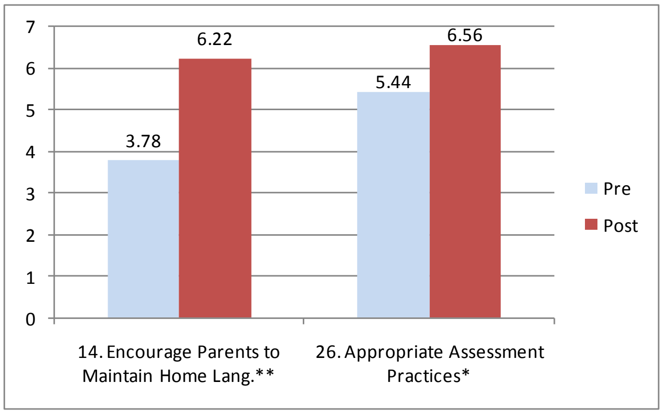
[9] The 26 CASEBA items are presented in their original order (numbered consecutively) – the items are not grouped exactly by scale within the tool – for example, please note that three items make up the *Culturally Responsive Instruction* scale – Items 3, 23, and 24.

Scale[9]	Item	Description	Pre	Post	Difference
Supports for English Language Development (continued)	19	Teaching staff interact one on one with all children in ways that support the acquisition of English.	5.33	5.78	0.45
	20	Teaching staff expand children's repertoire of concepts and vocabulary in English.	3.67	3.33	-0.34
Supports for Literacy Development in English	21	Books, print, and literacy props are available in English.	5.11	5.22	0.11
	22	Teaching staff support the learning of print-related early literacy skills in English.	3.33	3.22	-0.11
Culturally Responsive Instruction	23	Teaching staff provide a warm, emotionally supportive, and low-anxiety classroom environment for all children.	6.56	6.22	-0.34
	24	Teaching staff foster a calm and respectful learning environment in which all children are able to hear adult talk.	6.56	6.44	-0.12
Supports for English Language Development	25	Teaching staff create a content-rich curriculum that offers meaningful opportunities to acquire and use new language skills.	4.11	4.56	0.45
Not Applicable	26	Teaching staff use appropriate assessment practices to identify children's language strengths and needs in their home language and in English.	5.44	6.56*	1.12

*p<.05, **p<.01, ***p<.001

Seven of the 26 items showed a significant change in their scores from pre- to post-observation. CASEBA items that showed statistically significant growth in scores are shown in Exhibit 13.

Exhibit 13. CASEBA Items with Statistically Significant Score Increases, from Pre- to Post- Observation



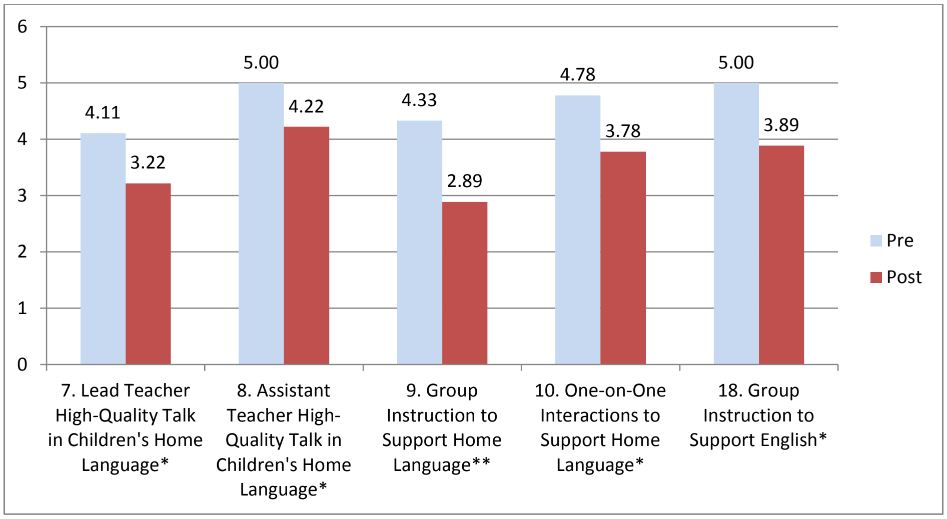
*p<.05, **p<.01

For Item 14 (*Teaching staff encourage DLL parents to maintain children's home language*), the average score increased significantly from 3.78 to 6.22 from the pre- to the post-observation. The post-observation average score of 6.22 (Good to Strong Evidence) indicates that teachers verbally encourage parents to use their home language at home and explain the value of home language for children's development, and that a lending library is available to parents with materials in their home language. In addition, some, but not all, indicators from the "Strong Evidence" rating are also met, which focus on encouraging parents verbally and in writing to talk, read, and write with children in their strongest language, sending home correspondence about the value of home language, and encouraging parents to participate in their child's program in their home language.

For Item 26 (*Teaching staff use appropriate assessment practices to identify children's language strengths and needs in their home language and in English*), scores increased significantly, from 5.44 to 6.56, with the post- score falling in the "Good to Strong Evidence" category. This item includes indicators related to the use of standardized tools to measure children's English and/or home language proficiency skills, use of assessment information by teachers for the purpose of communicating with parents, availability of bilingual staff, information sharing among staff about children's needs and progress, and observation of children to determine children's strengths in home language and proficiency in English.

Five items showed statistically significant decreases from the pre- to the post-observation, as shown in Exhibit 14.

Exhibit 14. CASEBA Items with Statistically Significant Score Decreases, from Pre- to Post-Observation



*p<.05, **p<.01

Four of the five items that showed a significant decrease from pre to post-observation are part of the *Supports for Home Language Development* scale. Item 7 (*The lead teacher uses high-quality talk in the children's home language*) and Item 8 (the same item, but for the assistant teacher) both showed a significant decrease: 0.89 points and 0.78 points respectively. Item 9 (*Teaching staff use effective strategies during group instruction to support DLL children's development of their home language*) also showed a significant decrease of 1.44 points. Item 10 (*Teaching staff interact one on one with DLL children in ways that support development of the home language*) decreased by 1.00 point from pre- to post-observation. There was a similar finding for Item 18, which falls under the *Supports for English Language Development* scale (*Teaching staff use effective strategies during group instruction to build children's communicative skills in English*), with a decrease of 1.11 points from the pre- to post-observation.

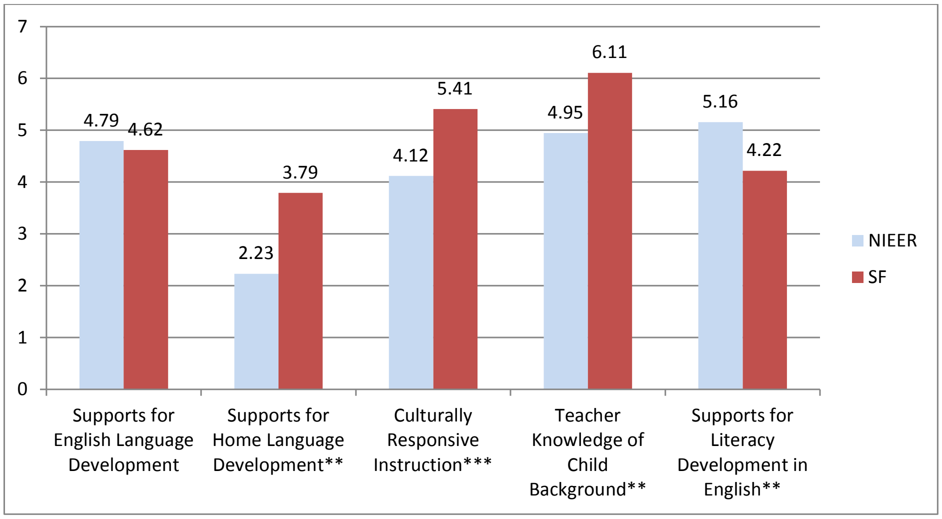
The post-scores for the four items under the *Supports for Home Language Development* scale were in the "minimal" or "minimal to good" evidence categories of the CASEBA rating scale. For example, in a typical classroom[10] among those observed, the lead teacher occasionally (as opposed to sometimes or often) used talk in one or more of the DLL children's home language that was lexically complex. There was "minimal evidence" of the use of effective strategies during group instruction to support home language development (e.g., teachers occasionally used the home language during book discussions or small-group activities, teachers engaged children in the singing of songs, chants, or finger plays in children's home language, or teachers tended to ask questions in the home languages that called for predetermined answers). There was "minimal to good evidence" that teaching staff used one-on-one interactions to support home language development—for example, teachers occasionally or sometimes (as opposed to often) initiated conversation in the home language or responded to individual children in the home language. In some classrooms, teachers sometimes asked DLL children open-ended questions or expanded on children's ideas or descriptions in their home language—but not often.

Comparison CASEBA Data

The post-observation San Francisco CASEBA scores are shown in Exhibit 15, in comparison to NIEER's validation study of the CASEBA in 120 public school, childcare, and Head Start classrooms in school districts with large Spanish bilingual populations (Freedson, Figueras-Daniel, Frede, Jung & Sideris, 2010)

[10] Evidence of specific indicators for each anchor rating (i.e., 1, 3, 5) varied across classrooms. This summary provides a picture of what a "typical" classroom might look like.

Exhibit 15. NIEER and San Francisco (Post-Observation) CASEBA Scale Scores



*p<.05, **p<.01, ***p<.001

Commented [GF2]: Same issue – is this spacing off?

The exhibit shows that San Francisco's CASEBA scale scores (post-observation) are relatively similar to those found in the NIEER study. San Francisco scores were higher than the classrooms in NIEER's study for three scales: *Supports for Home Language Development*, *Culturally Responsive Instruction*, and *Teacher Knowledge of Child Background*. *Supports for Home Language Development* received the lowest score among the five scales in both studies, although San Francisco classrooms scored higher than those in NIEER's study (3.79 compared to 2.23). *Supports for English Language Development* scores in the NIEER and San Francisco were about the same (4.79 and 4.62 respectively). The average *Supports for Literacy Development in English* score was higher in the NIEER study than in San Francisco (5.16 compared to 4.22). All of the differences between the NIEER and San Francisco CASEBA scales scores are statistically significant, except for *Supports for English Language Development*.

Correlations

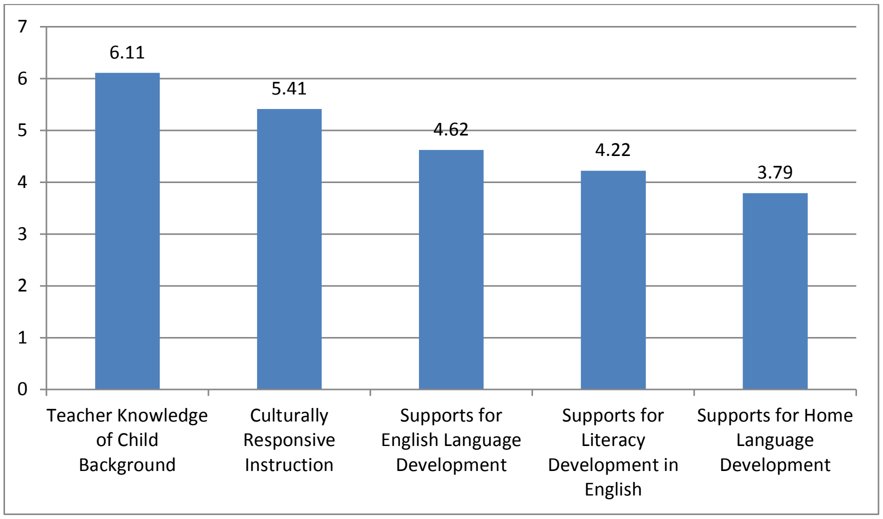
Analyses were conducted to identify whether the post-observation scores for the five CASEBA scales were positively or negatively correlated with each other. The *Supports for Home Language Development* scale was significantly positively correlated with *Culturally Responsive Instruction*. That is, classrooms that scored high on one scale also scored high on the other (or if they scored low on one scale, they scored low on the other). *Teacher Knowledge of Child Background* was significantly negatively correlated with *Supports for English Language Development* and with *Supports for Literacy Development in English* – meaning when a classroom scored high on one scale, they scored low on the other (and vice versa). The other post-intervention scale scores were not statistically significantly correlated with each other.

Analyses were also conducted to identify whether score changes, from pre- to post-intervention, for the five CASEBA scales were correlated with each other. Changes in the *Supports for English Language Development* scale were significantly positively correlated with changes in the *Supports for Literacy Development in English* scale. That is, teachers who had positive gains on one scale were likely to have positive gains on the other. None of the other changes in score, from pre- to post-intervention, were statistically significantly correlated with each other.

Conclusion

With the exception of one of the five CASEBA scales (*Teacher Knowledge of Child Background*), CASEBA pre- and post- scores did not significantly increase for the training participants. This may be a function of several factors – the five month period between the pre- and post-observation may not have been long enough for teachers to adopt new practices, based on the training; the tool may not be sensitive to changes that do occur within such a time period, and/or the training did not align well with the CASEBA tool. This latter factor is most likely, as the training developers described the intervention as focusing on a few specific strategies, whereas the CASEBA is a more far-reaching tool related to program practices, classroom resources, as well as teacher behavior. In general, the CASEBA pilot test provided a snapshot of language and literacy practices to support DLL children at two points in time –and offered First 5 San Francisco some baseline information about how teachers and preschool settings support dual language learners. In addition, the San Francisco post-CASEBA results are fairly consistent with the NIEER CASEBA findings.

Exhibit 16. Post- Observation CASEBA Scale Scores



In sum, the post-CASEBA observation scale scores indicated that:

- Programs and teachers demonstrated that they know the cultural and linguistic backgrounds of the children they serve (average score of 6.11).
- Classrooms also scored highly (average score of 5.41) in regard to incorporating the cultural backgrounds and life experiences of DLL children in the classroom, and in providing an emotionally warm and respectful environment for children.
- Classrooms scored in the mid range of the CASEBA rating scale (an average score of 4.62) in regard to teachers' use of high-quality talk in English and effective strategies to scaffold children's comprehension of instructional content in English.
- Classrooms scored similarly (average score of 4.22) in regard to the support of DLL's print literacy in English—this factor included indicators related to books, print, and literacy props in English and supporting children in learning print-related early literacy skills in English.
- The lowest score among the five scales was for *Supports for Home Language Development*, which received an average rating of 3.79. This factor includes a total of 11 items on the CASEBA, which relate to the use of home language for instructional purposes, the use of children's home language despite teachers' proficiency in the language, the use of high-quality talk in children's home language, and effective strategies to support children's development of their home language. In addition, the availability of books, print, and literacy props; teachers' support of the learning of print-related early literacy skills in the DLL children's home languages; and teachers' encouragement of DLL parents to maintain children's home language are included in this scale.

In general, AIR observers reported that the CASEBA was relatively easy to use and that they were able to observe and score items with confidence. The CASEBA is designed to be used in any preschool setting, regardless of the linguistic or cultural background of the children served. The tool was validated by NIEER in classrooms with a high proportion of Spanish-speaking DLL children. AIR staff generally felt the tool mostly worked well in both types of classrooms they observed for this study (with Spanish-speaking and Cantonese-speaking DLLs). If First 5 San Francisco were to use the tool in the future, however, AIR recommends that they work in partnership with NIEER to refine the CASEBA training and observer guide to include discussion of strong examples of best practices in PFA's Cantonese-dominant classrooms. Training serves several purposes—including identifying key examples observers may encounter in classrooms serving different cultural and linguistic groups and helping trainees to articulate their own assumptions and how they might influence their use of the tool. The small number of AIR observers for this study—three were used—enabled them to meet frequently to discuss their observations and seek clarification from NIEER on some items during the process.

In sum, the CASEBA provides rich detail on strategies and supports for DLL children in preschool settings, yet its format enables staff to administer the tool easily and within one observation session. In addition, the CASEBA can serve as a springboard for professional development efforts for teaching staff—most of the tool's items and indicators are easily understood and can inform training efforts.

References

Burchinal, M. R., Peisner-Feinburg, E., Pianta, R. C., & Howes, C. (2002). Development of Academic Skills from Preschool through Second Grade: Family and Classroom Predictors of Developmental Trajectories. *Journal of School Psychology, 40*(5) 415-36.

Freedson, M., Figueras, A., & Frede, E. Classroom Assessment of Supports for Emergent Bilingual Acquisition (2008). New Brunswick, New Jersey: National Institute for Early Education Research, Rutgers, The State University of New Jersey.

Freedson, M., Figueras-Daniel, A., Frede, E., Jung, K., & Sideris, J. (2010, June). *The Classroom Assessment of Supports for Emergent Bilingual Acquisition (CASEBA): Validating an observational rating scale of preschool classroom quality for dual language learners*. Poster presented at Head Start's 10th National Research Conference, Washington, DC.

Hamre, B. K., & Pianta, R. C. (2005). Can Instructional and Emotional Support in the First-Grade Classroom Make a Difference for Children at Risk of School Failure? *Child Development, 76*(5), 949-967.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs. *Early Childhood Research Quarterly, (23)*1, pp. 27–50.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development, (79)*3, pp. 732–749.

Mashburn, A. J., & Pianta, R. C. (2006). Social Relationships and School Readiness. *Early Education and Development, 17*(1), 151-176.

Mitchell-Copeland, J., Denham, S. A., & DeMulder, E. K. (1997). Q-sort assessment of child-teacher attachment relationships and social competence in the preschool. *Early Education and Development, 8*, 27-39.

Peisner-Feinberg, E.S., Burchinal, M.R., Clifford, R.M., Culkin, M., Howes, C., Kagan, S.L., Yazejian, N., Byler, P., & Rustici, J. (1999). *The Children of the Cost, Quality & Outcomes Study Go to School: Technical Report.* Chapel Hill, NC: Frank Porter Graham Child Development Center, UNC-Chapel Hill.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) Manual.* Baltimore, MD: Brookes Publishing.