Study of the Impact of the California Formative Assessment and Support System for Teachers: Research Summary

Marnie Thompson, Pamela Paek, Laura Goe, and Eva Ponte

July 2004

The California Formative Assessment and Support System for Teachers (CFASST) is a structured, two-year induction program for beginning teachers that is used across the state as a central component of California's Beginning Teacher Support and Assessment (BTSA) program. In 2002, the California Commission on Teacher Credentialing commissioned ETS to conduct a study of the impact of BTSA/CFASST. In this study, we looked at the impact of the program on the teaching practices of beginning teachers and on the learning of their students. This document summarizes the results of that study. Detailed information on the methods and findings of the study can be found in a series of technical reports that can be obtained from www.ets.org/research/CFASST.html:

Report 1: Beginning Teachers' Engagement with BTSA/CFASST
Report 2: Relationship of BTSA/CFASST and Teacher Practices
Report 3: Relationship of BTSA/CFASST and Student Achievement
Report 4: Methodological Considerations and Recommendations for Future Research

Overview of BTSA/CFASST and the Design of the Study

CFASST provides a series of twelve "events" based on the California Standards for the Teaching Profession (CSTP). With the guidance of an experienced teacher who has been trained as a support provider, beginning teachers learn about best practices, plan lessons, reflect on their teaching, and apply what they have learned in their classrooms. This is facilitated by ongoing formative assessment in which beginning teachers and their support providers assess their teaching practice and set goals for professional growth, using a formative assessment tool based on the CSTP. In the 2002-2003 school year, CFASST was in its fourth year of large-scale use. In that year, 133 of 142 BTSA programs employed CFASST as a central component.

We employed a quasi-experimental design to assess the impact of BTSA/CFASST on teacher practices and student learning. The study sample was drawn from the population of grade 3 to 5 teachers who were in their third year of teaching in the 2002-2003 school year and who had participated in BTSA in their first two years of teaching. Because BTSA is a statewide program for all new teachers with Multiple or Single Subject preliminary credentials, there was no natural control group by which to estimate program impacts. To identify a comparison group for the study, we capitalized on the fact that implementation of BTSA and CFASST is variable, such that some BTSA enrollees engage with the program in a deep and sustained way, while others receive far less of the "treatment." A considerable portion of engagement variability stems from program-level differences in such matters as ease and frequency of access to support providers; time to meet and work on the CFASST events; and training and monitoring of support providers. We identified a sample of BTSA graduates who had a high level of engagement with the program and compared them to a sample that had low or no engagement.

We began the study with a survey of 1,125 third year teachers that asked teachers about their experiences with BTSA and CFASST. The survey was completed by 287 teachers (a 26% response rate), from 78 BTSA programs. From teachers' responses, we calculated a CFASST engagement score by which we classified each teacher into high, middle, or low CFASST engagement levels. We contacted a sub-sample of 64 respondents for further study, attempting to draw from the top and bottom of the CFASST engagement scale. These teachers were interviewed by phone for further information and to validate the survey results. From this group, we recruited 34 teachers for blind case studies involving classroom observations and face-to-face interviews.

To estimate impacts on teacher practices, we developed ratings on ten measures of teaching practice that have been empirically or theoretically linked with effective teaching: instructional planning, reflection on practice, questioning (three measures), feedback (three measures), depth of student understanding, and overall teaching

practice (CSTP score). We hypothesized that the higher a teacher's CFASST engagement level, the stronger their teaching practices would be.

To estimate the impact of BTSA/CFASST on student learning, we collected data from California's Standardized Testing and Reporting (STAR) program for the students of 144 survey respondents. We hypothesized that the higher a teacher's CFASST engagement level, the better their students would do on the STAR tests. To account for effects contributed at the school level, we employed a school-level performance indicator, the Academic Performance Index (API), a numeric score calculated by the State that reflects a school's performance in the STAR program. Using API as a covariate, we can be reasonably assured that differences in outcomes for the students of high and low CFASST engagement teachers are not the result of differences between schools.

Teachers' Experiences with BTSA/CFASST

Most teachers had some form of CFASST orientation (all but 5%) and a support provider (except for 7% the first year and 3% the second year); numbers that, though low, represent a challenge to the core of BTSA. The majority of teachers (more than 70%) thought that their support providers were "warm and supportive," and trusted them to the point of feeling comfortable enough to "share everything" with them. Three-fifths of beginning teachers had support providers who were located in the same school; two-fifths had remote support providers. There was high variability in the frequency of meetings between the beginning teachers and their support providers, ranging from less than once a month to more than once a week. For most teachers, the primary focus of the meetings was on emotional support, with instructional support and support for managing student behavior coming just behind. Teachers reported completing CFASST events at varying rates. Interview data revealed that having a support provider was viewed as the best part of being in BTSA, followed by the CFASST events themselves. Interview data also indicated a strong relationship between CFASST engagement and having a support provider on-site. Interviews also pointed to some complaints about the program, the primary one being the "paperwork" required, along with the program being repetitive and taking too much time. A cross-analysis of the survey and interview data confirmed the CFASST engagement levels generated from the survey: 29% low engagement, 45% mid-level engagement, and 26% high engagement. There was a small but statistically significant correlation between CFASST engagement and school API score ($r = .14$, $p < .05$).

BTSA/CFASST Engagement and Teacher Practices

Comparing the high and low CFASST engagement groups with regard to teaching practices (Table 1), we found one measure that showed a statistically significant difference: Instructional Planning. For seven of the ten measures of teaching practice, the high engagement group had a greater mean score than the low engagement group: Instructional Planning, Reflection on Practice, Proportion of Questions that are Intermediate, Proportion of Questions that are Open, Proportion of Feedback that is Positive, Proportion of Feedback that is Substantive and Specific, and Depth of Student Understanding. On the three remaining measures (Proportion of Questions that are Deep, Proportion of Feedback that is Instructional, and CSTP Score) the score differences were close to zero.

Table 1. Mean Scores and Effect Size of Measures of Teaching Practice, by CFASST Engagement Level

Measure of Teaching Practice	Low CFASST Engagement (*N*=12)			High CFASST Engagement (*N*=15)			Effect Size*
	Mean	SD	Range	Mean	SD	Range	
Instructional Planning	17.3	1.4	16-20	19.5	1.6	17-22	1.46
Reflection on Practice	31.9	4.1	25-38	33.1	3.0	28-40	0.33
% Deep Questions	2.2	2.9	0-8	2.0	3.3	0-11	-0.06
% Intermediate Questions	29.1	17.4	6-59	34.7	20.0	5-68	0.30
% Open Questions	27.0	21.1	0-66	32.2	20.9	0-76	0.25
% Positive Feedback	82.2	16.4	36-98	85.4	11.0	62-100	0.23
% Instructional Feedback	85.5	14.9	7-98	84.9	12.7	56-100	-0.04
% Substantive/Specific Feedback	15.0	8.9	2-32	19.9	9.4	9-44	0.54
Depth of Student Understanding	1.6	0.4	1.1-2.5	1.7	0.4	1.2-2.4	0.25
CSTP Score	56.1	12.5	34-70	55.7	10.1	38-74	-0.04
						Mean Effect Size	0.32

* Cohen's *d*

If BTSA/CFASST had no effect, we would expect the low CFASST group means to be larger than the high CFASST group means on roughly half the measures, since the correlations among the measures was not high. The fact that this did not occur suggests that BTSA/CFASST has a positive impact on teaching practices. These generally positive patterns were reflected in the effect sizes for the ten measures of teaching practice. These range from near zero to 1.46 standard deviations (SDs), with most in the range of 0.23 to 0.54. The average effect size is 0.32 SDs. It is important to remember that we are talking about effect size with regard to teaching practices, which does not translate directly into equivalent effects on student outcomes. The impact on students of a change in teacher practice depends on the nature, frequency, and centrality of the teaching practice.

BTSA/CFASST Engagement and Student Learning

In comparing the performance of the high and low CFASST engagement groups with regard to student learning, we found a similar pattern of consistently positive differences in favor of the high CFASST engagement teachers, even after controlling for API score. Though no tests showed statistically significant differences, effect sizes ranged from 0.03 to 0.40 standard deviations, with an average effect size of 0.25 SDs (Table 2). The greatest effect size was for the CAT-6 Math test (0.40 SDs), with most in the .20s and .30s. To put these effect sizes in context, it is useful to consider that the recent re-norming of the SAT-9 achievement test showed differences between 4th and 5th graders to be around 0.50 SDs in math and 0.33 SDs in language arts. Thus, the average effect size found here—0.25 SDs—might be seen as equivalent to half a year's growth or more.

Table 2. Mean Scores and Effect Size of STAR Test Scores, by CFASST Engagement Level

STAR Test	Low CFASST Engagement (*N*=28)			High CFASST Engagement (*N*=45)			Effect Size*
	Mean**	SD	Range	Mean**	SD	Range	
CAT-6 Reading	40.00	8.73	13-67	42.04	8.63	16-68	0.23
CAT-6 Language	42.25	8.89	15-69	43.89	8.79	17-71	0.19
CAT-6 Spelling	47.94	8.67	21-74	51.08	8.58	25-77	0.36
CAT-6 Math	48.43	9.66	19-78	52.23	9.56	23-81	0.40
CST ELA	3.02	0.38	1-5	3.03	0.38	1-5	0.03
CST Math	2.92	0.75	1-5	3.13	0.74	1-5	0.28
						Mean Effect Size	0.25

* Cohen's *d* ** Adjusted for API

Effect of BTSA/CFASST Engagement Relative to Student and School Effects

We employed hierarchical linear modeling (HLM) to look at student-level variables nested within individual teachers' classrooms. Table 3 reveals the contributions of teachers' engagement with BTSA/CFASST, API score, and three student-level variables that have been shown to be negatively correlated with achievement: having a disability designation, low socio-economic status (or low SES, represented by free or reduced lunch status), and being identified as an English language learner (ELL).

It can readily be seen in Table 3 that the student-level variables have a consistently negative relationship with test score, whereas CFASST engagement level has a consistently positive relationship. (None of the CFASST coefficients are statistically significant.) API score has a positive relationship with student scores on the CAT-6 tests, but shows no relationship with scores on the California Standards Tests.

Table 3. HLM Analysis: Coefficients for School, Teacher, and Student-level Variables for STAR Tests (*N* = 115)

Coefficients		CAT-6 Reading	CAT-6 Language	CAT-6 Spelling	CAT-6 Math	CST ELA	CST Math
School Effects (API)		0.10	0.11	0.09	0.11	0.00	0.00
Teacher Effects (CFASST Engagement)		2.04	1.48	2.08	2.06	0.02	0.12
Student Effects:	ELL Status	-10.77	-7.86	-5.12	-5.32	-0.39	-0.14
	Low SES Status	-9.82	-10.00	-5.82	-9.15	-0.41	-0.38
	Disability Status	-16.48	-17.75	-18.72	-18.80	-0.73	-0.73

On the various CAT-6 tests, students with disabilities scored 16 to 19 points lower, on average, than those without a disability. ELL students scored 5 to 11 points lower, and low SES students scored 6 to10 points lower, on average, than students without those designations. Raising the achievement of students in these groups has become a priority for most schools, so it is useful to see how much difference CFASST engagement makes in accomplishing that goal. Table 3 shows that for each successive CFASST engagement level (low to middle or middle to high), students perform 1.5 to 2 percentiles higher on CAT-6 tests. In other words, students of teachers with a high level of CFASST engagement score 3 to 4 points more, on average, than students of teachers with low CFASST engagement. Additional analyses confirm that this positive effect holds within all three sub-groups. BTSA/CFASST, by itself, does not ameliorate these correlates of lower achievement, but it does counteract them.

Putting the Results in Context

Overall, our findings show a positive impact of BTSA/CFASST on teachers and students. The fact that there were positive effects both for teachers and for students is especially encouraging, as it supports our model of how BTSA/CFASST works: the support of an experienced teacher, the curriculum of CFASST events, and the formative assessment aspects of BTSA/CFASST combine to improve beginning teachers' practices. These improved practices, in turn, lead to improved student learning, equivalent to half a year's growth or more. The results of this study have relevance to other mentor-based induction programs, to the degree that such programs have similar components and are implemented so that teachers have a high level of sustained engagement.

It is important to consider the limitations of this study when weighing these results. In particular, there are two issues that arise from the fact that we were unable to use a random assignment design. The use of the low CFASST engagement group in the place of a true "control" reduces the magnitude of the contrast between the groups being compared. Thus, our effect sizes and significance statistics are likely to be underestimates. On the other hand, our use of a retrospective quasi-experimental design means that we cannot definitively attribute all measured effects to the treatment. That is, there may be unmeasured characteristics of the schools, teachers, or students that account for some of the differences in performance between the high and low engagement groups. A randomized experiment or a quasi-experimental study with additional control variables will be needed to address this issue. It will also be important to increase the number of teachers in any future study, to enhance the likelihood of detecting significant differences between "treated" and "untreated" teachers and their students.

Despite these limitations, this study makes an important contribution to the research base on new teacher induction. The "logic" of such programs has been well articulated in the induction literature and documented in qualitative studies and small sample pre-post studies. But the field is only now accumulating quantitative studies that show measurable effects on teacher retention, teacher practices, and student learning. With regard to retention, Richard Ingersoll's longitudinal analyses show that high quality induction can cut new teacher turnover by more than half. This study makes an empirical connection between induction, changes in teacher practices, and improved student learning, addressing another gap in the research base. Such studies have proven challenging due to difficulties in identifying comparison groups, in reliably knowing what teachers are doing in their classrooms, and in obtaining student achievement data traceable to specific teachers. Given these challenges and the paucity of research making these connections, this study represents an important result in the chain of evidence surrounding the effectiveness of new teacher induction.

These findings are timely given the emphasis placed by the No Child Left Behind Act on the professional development of teachers, with the expectation that improvements in professional development will promote positive changes in teaching practices, which will in turn enhance student achievement. Many states and districts are now establishing guidelines and funding streams for new teacher induction, and this study can increase their confidence in the efficacy of such programs. With high quality induction costing $2,500-$4,000 per teacher, it is important to be able to quantify its benefits. The cost savings stemming from reduced teacher turnover can offset much of the cost of induction. The benefits to teaching and learning, as documented in this study, can certainly be seen to justify the remainder of the expense. Thus, policymakers now have an enhanced empirical basis for supporting high quality teacher induction programs. However, the conditions that made this study possible—high variability in the strength of new teachers' exposure to the program—are not conducive to the improvements that educators are seeking. Induction programs will only have widespread positive effects when they are implemented so that all new teachers have a high level of sustained engagement in substantive professional learning.