

Writer CoachConnection™

Berkeley Unified School District Writing Assessment Analysis, 2007-08

October 2, 2008

Part I

Executive Summary

Robert Menzimer, Executive Director, Community Alliance for Learning

It is difficult to isolate and evaluate factors contributing to educational achievement, yet policy decisions still have to be made in times of limited budgets, necessitating efforts to answer the question, “Is this program effective?”

Toward that end, independent evaluation consultant PJ Hallam, Ph.D., worked with WriterCoach Connection (WCC) to conduct a limited causal study based on Berkeley Unified School District fall/spring 7th and 8th grade writing assessment results. Dr. Hallam evaluated the effectiveness of writer coaching on student achievement using these timed writing tests as one indicator of overall development of writing skills. This study looked at the 7th grade scores at one middle school (Longfellow) where students received writer coaching, and compared their scores against a similar middle school (Willard) where students did not receive individual writer coaching. In addition, the researcher analyzed the overall performance of 8th graders at all Berkeley Unified middle schools, since all 8th graders participate in the writer coaching program. Within this group, the performance of 8th grade students who had two years of writer coaching (Longfellow) was compared with the remaining 8th graders who had only one year of writer coaching.

The findings are heartening. An executive summary is presented here, followed by the 7th and 8th grade studies below. *Overall, student writing at all levels improved, a testament to the tremendous efforts the district has put toward writing instruction in recent years.* With regard to the possible impact of writer coaching on that student improvement, these are the key findings:

1. Students who had the benefit of writer coaching in 7th grade (Longfellow) performed significantly higher than 7th graders who did not participate in the writer coaching program (Willard). When asked to write an on-demand essay for a new topic in the spring writing sample, a higher number of coached students were able to successfully demonstrate control of the more complex aspects than were non-coached students, and fewer coached students who struggle with writing found it as difficult as did non-coached students. These results indicate that coaching helped students at both the higher and lower ends of the writing-ability spectrum. Combined with the WCC survey results (in a separate report, at www.writercoachconnection.org), which strongly indicate that both teachers and students thought that coaches helped 7th grade students with writing skills, these findings provide

evidence that supports the contention that **WCC is a contributing factor to Longfellow's stronger performance on the BUSD writing assessment.**

2. **BUSD 8th grade writing scores were statistically significantly higher in the spring than in the fall.** Longfellow, where 8th graders had individual attention from writer coaches for two consecutive years, saw student scores increase by almost a full point (0.76). Although it's not possible to isolate WCC as a factor, the startlingly high increase at Longfellow is a measurement of writing skills in students who were completing a second consecutive year of writer coaching, as compared to their counterparts at King and Willard, who showed improvement but at a somewhat lower level, with just a single year of coaching. In addition, 8th grade teachers' and students' survey responses indicate that they find WCC to be very helpful.

Although improvement in writing and critical thinking skills cannot be attributed solely to WCC in a landscape as complex as the teaching of English Language Arts, **the results of the BUSD-based writing assessment are very encouraging in terms of what seems to be the impact of writer coaching on Berkeley middle school writing.** The collaboration of WCC in support of the many elements of effective teaching and learning appears to play a significant role in:

1. Higher levels of writing achievement among 7th graders who have had writer coaching as compared to 7th graders who have not been coached,
2. Higher levels of writing achievement among eighth graders who have had two years of writer coaching as compared to eighth graders with only one year of coaching,
3. Increased levels of writing achievement across the spectrum of writing abilities for both 7th and 8th graders, indicating that coaching all students is an effective approach, and
4. The overwhelmingly positive survey-based program assessment by all program participants: students, teachers, and coaches (separate report on website at www.WriterCoachConnection.org).

These results support what we know in our hearts, after eight years of experience, to be true: WriterCoach Connection works.

One additional note:

Community Alliance for Learning (CAFL) is honored to collaborate with BUSD efforts toward improving student writing, was pleased to provide volunteer scorers for the spring 2008 reading, and will be happy to continue to do so for future assessment scorings. It is noteworthy that the statistical IRR (interrelater reliability) for the spring scoring was higher than the fall scoring, perhaps indicating that the presence of WCC volunteers both contributed to a lessening of reader fatigue and placed heightened attention to consistency regarding anchor papers and scoring standards. CAFL looks forward to continuing this collaboration.

Part II

Seventh Grade Longfellow and Willard Writing Sample Analysis, 2007- 08

PJ Hallam, Ph.D., Independent Evaluation Consultant

Teresa Barnett, Program Manager, Community Alliance for Learning

In school year 2007-08, WriterCoach Connection provided coaches for all 7th grade students at Longfellow Middle School, while Willard 7th graders did not participate in the coaching program. Since these two middle schools are similar in student populations and had similar fall test-score results, this report compares their spring scoring results. Its findings document that **Longfellow 7th graders' spring scores were significantly higher, while Willard students' scores were not**, and explores possible contributing reasons for this discrepancy.

Seventh Grade Background

In school year 2007-08, 7th graders in BUSD completed two “on-demand,” timed essay assessments, both in the persuasive genre. In the fall, they wrote to the topic of whether the schools should switch to a year-round schedule. In the spring, they wrote to the topic of whether children under the age of 16 should be allowed to work. (Prompt material available on request.)

Both sets of essays were scored using the traditional process of first calibrating readers' interpretations of the four-point scoring rubric, based on the California state rubric for persuasive writing, by having readers score exemplars prior to scoring the students' anonymized papers twice. A final score was compiled by combining the two readers' scores, creating a range of two to eight points. (For more information about the scoring process, see Rebecca Cheung, Director of Evaluation and Assessment, Berkeley Unified School District.)

Seventh Grade Student Improvement over Time

Table 1 shows the mean fall and spring writing scores of all the 7th grade students at Longfellow and Willard. Not surprisingly, students' spring mean scores are higher than the fall mean scores, as students learn more about writing skills over the school year.

Table 1 Longfellow and Willard 7th Grade Mean Scores

	Fall Mean	Spring Mean	Difference	P-Value*	Significant	<i>N Fall or Spring</i>	<i>N Fall and Spring</i>
Longfellow	4.27	4.52	0.25	0.03	yes	129	126
Willard	4.16	4.19	0.02	0.84	no	147	129

* Matched pair test of significance

The increase in Longfellow 7th graders' mean fall and spring scores (0.25) is statistically significant (< .05 level, meaning the substantial increase in mean scores is highly unlikely due to chance) while Willard students' increase (0.02) is not statistically significant (the increase is small, and chance may have played a part in it).

Seventh Grade Interrater Reliability

A major contributing factor to stakeholders' confidence in scores based on human judgment is interrater reliability. The interrater correlation for fall was 0.69, and for spring it was 0.68 (Pearson). These levels of interrater reliability are considered to be moderately reliable. Coefficients higher than .75 are more ideal, but stakeholders, such as teachers, principals, parents, and community members, can have moderate confidence that the differences between fall and spring scores is due to student differences in ability, not to variability of scorers' interpretation of the rubrics.

Seventh Grade Student Scoring Patterns

Displaying the students' fall and spring scores in crosstab table format in Tables 2 and 3 provides possible insights into students' scoring patterns at the two schools.

Table 2: Longfellow Fall and Spring Scores Crosstab

Longfellow		Spring Score							Grand Total	
Fall Score		0	2	3	4	5	6	7		8
0		1								1
2			1	2	9					12
3			2	6	9	1	3	1		22
4			1	7	25	10	5			48
5			1		11	2	2	1	1	18
6					3		8	2		13
7						1	4	2	1	8
8					1		1	1	1	4
Grand Total		0	6	15	58	14	23	7	3	126

Total students with higher scores in spring: 48 (Above shaded cells)

Total students with lower scores in spring: 43 (Below shaded cells)

Table 3: Willard Fall and Spring Scores Crosstab

Willard		Spring Score							Grand Total	
Fall Score		0	2	3	4	5	6	7		8
0		1			1					1
2			7	6	5					18
3			3	7	8	1		1		20
4			4	10	20	10	1	1		46
5				1	10	5	4	3		23
6					3	1	4	1		9
7					2	3	2	1		8
8						1	2	1	1	4
Grand Total			14	24	49	21	9	11	1	129

Total students with higher scores in spring: 46 (Above shaded cells)

Total students with lower scores in spring: 47 (Below shaded cells)

Understanding Crosstab Tables

In crosstab tables, the rows display fall scores and the columns show where these students scored in the spring. For example, starting at the top left of Table 2, the first row displays one fall score of zero. In the spring, this student scored a 2, so it is under the 2 column. The next row displays fall scores of 2. One student who scored a 2 in the fall also scored a 2 in the spring. Note that the number of students who scored the same in fall and spring are indicated by shaded cells. Two of the students who scored 2 in the fall increased their score to a 3 in the spring, and 9 increased to a 4. Note that these improved scores are located above the shaded cells in the table. The next row displays fall scores of 3. Two of these students scored a 2 in spring, lower than in the fall. Note that students who are lower in the spring are below the shaded cells.

Analysis of Crosstab Tables 2 and 3

The number of students with higher spring scores than fall scores at Longfellow was 48 (38%); at Willard it was 43 (36%). The number of students with lower scores at Longfellow was 43 (34%); at Willard it was 47 (36%). The remaining students at both schools scored the same in fall and spring. These findings indicate that a few more students at Longfellow improved and a few less students at Longfellow decreased than did students at Willard.

Dividing the scores into two groups, the lower scoring range, 0 to 5, and higher range, 6 to 8, provides more informative insights. According to the scoring rubric descriptors, students who score at the higher levels of 6 to 8 (levels 3 and 4 for single scores) display more control of the writing process and are more likely to be able to handle regular classroom writing assignments.

In the fall, the number of Longfellow 7th graders who scored in the lower range was 91, and higher range was 25. In the spring, 33 were in the higher range, indicating that 8 students improved. At Willard, the fall numbers were 108 and 21 in *both* fall and spring. More Longfellow students moved to the higher scoring range.

At both schools, student's scores dropped at the 2, 3, 4 and 5 levels in the spring as students tried their hands at the new topic on one of the hottest spring days in Berkeley's history. The majority of students dropped by one score point (adjacent and below the shaded cell), Longfellow N = 20, Willard, N = 23. However, Longfellow had half as many students drop lower by two or more points, N = 7, than did Willard, N = 14. Apparently fewer Longfellow students struggled deeply with the new topic than did Willard students.

Seventh Grade Conclusions

The data indicate that when asked to write an on-demand essay for a new topic in the spring writing sample, students with coaching had higher mean scores than those who did not. Also, more Longfellow students were able to successfully demonstrate control of the more

complex aspects than were Willard students, and fewer Longfellow students who struggle with writing found it as difficult as did Willard students.

In terms of closing the achievement gap, it is noteworthy that this evidence indicates that **WriterCoach Connection helped students improve at both the lower and higher levels of writing ability.** While this investigation did not disaggregate by ethnicity, previous studies of BUSD student achievement disaggregated by ethnicity and SES indicate achievement scores are highly correlated with these demographic categories.

The underlying reasons for these discrepancies are likely due to a number of factors. This small, modest analysis is not a comprehensive investigation of casualty. However, the significantly higher performance of Longfellow students who had coaching compared to the performance of those who did not have coaching, combined with the WCC survey results from both teachers and students that strongly indicated coaches helped 7th grade students with writing skills specifically related to these types of writing improvements, provides some evidence to support the hypothesis that **WCC could be a contributing factor to Longfellow's stronger performance on the BUSD writing assessment.**

Part III

Eighth Grade Writing Sample and CST ELA Analysis, 2007-08

PJ Hallam, Ph.D., Independent Evaluation Consultant

Teresa Barnett, Program Manager, Community Alliance for Learning

In school year 2007-08, WriterCoach Connection provided coaches for all 8th grade students at BUSD's three middle schools. The findings in this report indicate that most **8th graders increased their writing scores significantly.**

Eighth Grade Background

In school year 2007-08, 536 8th graders in BUSD completed two "on-demand," timed essay assessments. Both writing prompts were in the persuasive genre. In the fall, students wrote on the topic of banning junk food in schools, and in the spring the topic was whether schools should conduct random drug testing. (Prompt material available on request.)

Both sets of essays were scored using the traditional process of first calibrating readers' interpretations of the four-point scoring rubric by scoring exemplars prior to scoring the students' anonymized papers twice. A final score was compiled by combining the two readers' scores. (For more information about the scoring process, see Rebecca Cheung, Director of Evaluation and Assessment, Berkeley Unified School District.)

BUSD 8th graders also completed two writing subsections of the California State Test (CST), writing strategies and writing conventions, which contributed to their overall ELA score, along with three reading subsections.

Eighth Grade Student Improvement in 2007-08 on BUSD Writing Assessment

While the fall and spring essays varied by topic, it may be conjectured that a construct of "general writing ability" underpins both, and thus improvement over time may be noted despite the complexities due to a switch in topics.

Table 1: SY 08 8th Graders' Fall and Spring Mean Scores on BUSD Writing Assessment

	Fall Mean	Spring Mean	Difference	P Value	Significant	N
All BUSD 8th Graders	4.24	4.70	0.46	<0.01	yes	536
King	4.31	4.72	0.40	<0.01	yes	275
Longfellow	4.22	4.98	0.76	<0.01	yes	140
Willard	4.10	4.32	0.22	0.06	no	121

Students at King and Longfellow had significantly higher 8th grade writing scores in the spring than in the fall. Longfellow 8th graders increased by almost a point (0.76), followed by King students with an increase of 0.40, and Willard students with an increase of 0.22, with a p-value of .06 (.05 is the significance cut point for social sciences). The students at

Longfellow received one-on-one writer coaching both in 7th and in 8th grade, and these results may indicate that **the long term individual attention on writing may have been a contributing factor in these students' success.**

Stakeholder Confidence in Human Judgment

The growth indicated by the BUSD 8th writing assessment is quite positive, and stakeholders such as teachers, parents, and administrators naturally would like to know to what extent this scoring data can be trusted. Two sources of evidence on the reliability of human judgment in scoring 8th graders' essays are provided below.

Eighth Grade Interrater Reliability

Interrater reliability looks at the extent to which the first scorer's decision matched the second scorer's decision. The interrater correlation for 8th grade scoring in 2007-08 for the fall was 0.68, and for spring it was 0.73 (Pearson). These levels of interrater reliability are considered to be moderately reliable. Coefficients higher than .75 are more ideal, but stakeholders can have moderate confidence that the differences between fall and spring scores is due to student differences in ability, not to variability of scorers' interpretation of the rubrics.

Correlation with CST ELA Scores

CST 8th grade ELA scores are not completely aligned with the BUSD assessments because students do not actually write for the CST; they answer multiple choice questions. BUSD assesses students with a more authentic writing assessment precisely for this reason. However, there are some constructs of writing covered by both types of tests. Correlating the scores from the assessment scored by humans with the scores from the assessment with answers determined in advance and scored by machines provides a type of reliability check. Assessments that were highly uncorrelated, or negatively correlated, would provide evidence that something was amiss.

Table 2: SY 08 8th Graders' CST and BUSD Writing Assessment Correlations

	Spring BUSD/Spring CST ELA Scores Correlation	Significant	N
All 8th Grade	0.67	yes	540
King	0.66	yes	282
Longfellow	0.66	yes	137
Willard	0.77	yes	121

The data in Table 2 indicate that BUSD's assessment scores are moderately correlated with CST scores, and thus provide another source of evidence for being reliable.

Eighth Grade Conclusions

The BUSD fall and spring writing scores indicate that 8th graders improved their writing skills over SY 2008, and that the scorers who provided these scores were reasonably

reliable. While there is not any direct evidence that WCC contributions helped with this increase, **student achievement levels were highest at the school at which students receive two years of writer coach support. In addition, teachers' and students' survey responses indicate that they find WCC to be very helpful.**