

## BAIP Pilot Study Data

Douglas Glasnapp

### *Introduction*

Following is a summary of several analyses based on the field-testing of the Blending Assessment with Instruction Program (BAIP) for 2008 and 2009. In this particular set of comparisons, the data were not adequate for grades 7, 8, and high school.

In 2008, initial investigations were conducted to address the impact of BAIP usage on achievement outcomes using quasi-experimental design approaches. Based on BAIP usage information available in the state of Kansas, three “level of use” groupings were made: non-BAIP users, low-BAIP users, and high-BAIP users. Achievement outcomes were examined using performance scores on the spring 2008 mandated state assessment tests, outcomes which were the instructional targets of BAIP lessons, and as a secondary measure, scores on the BAIP tutorial items themselves. State-mandated test scores from the spring of 2007 were used as covariates to control for differences in prior achievement across the three quasi-experimental BAIP usage comparisons groups. Analyses were conducted separately for students at grades 4, 5 and 6.

In total, 60 different analyses were conducted across the different grades and achievement outcome variables. In all analyses, students’ prior year 2007 state mandated test scores were used as a statistical covariate to adjust 2008 test scores used as outcome variables and control for any extraneous differences that might be attributable to existing differences in group ability. Samples sizes were large for all analyses conducted with none less than 100 students per BAIP usage group and the majority having 300 or more student data points per group. To summarize the results, all analyses detected statistically

significant results among groups ( $p$ -value  $< .001$ ). The statistically significant results were primarily due to the high statistical power in testing the group difference hypothesis as the sample sizes were extremely large for groups in the analyses. Of more importance than the statistical significance test result was the consistent pattern in the means, showing a trend of increasingly larger means across the BAIP usage groups with greater BAIP usage groups having the higher mean score. Effects reported as the difference in the “adjusted mean scores” between a BAIP usage group and the non-BAIP use comparison/control group indicated the extent to which a BAIP usage group scored higher than the non-BAIP user group on the achievement measure analyzed. Of the 108 “effect size” values reported, only 7 were in the negative range and reported as no effect. It is noted that while the majority of the evidence reported supports the conclusion that use of BAIP enhances student achievement, the results from the grade 6 analyses were strongest, followed by those at grade 4. The grade 5 evidence was also supportive of BAIP effectiveness.

In 2009, additional investigations were conducted to address the impact of BAIP usage on achievement outcomes. Data from three separate data bases, 2009 student tutorial usage data, 2009 teacher lesson use survey data and 2007 and 2009 student state assessment data, were combined to link BAIP usage to student performance on the state assessment measured outcomes in mathematics. Data were combined to support analyses at the student and building levels.

For these 2009 data, the number of tutorials taken by a student was an objectively determined measure and thus was highly trustworthy as an indicator of BAIP usage level. The teacher usage data were linked to students only if students in a teacher’s class took

tutorials. The collection of trustworthy data on teacher lesson use and exposure of students to lessons in a systematic manner is critical to the evaluation of overall BAIP effectiveness and is a primary goal of the proposed initiative.

Based on the number of tutorials taken only as a measure of BAIP usage, the data supporting BAIP effectiveness were encouraging for students at grades 5 and 6. As expected, students in the state with no evidence of BAIP usage showed no gain with z-score means approximately at zero for the 2007 and 2009 data (grade 5 2007 mean of .021 and 2009 mean of .017; grade 6 2007 mean of .020 and 2009 mean of .019). In comparison, however, students with evidence of having taken 15 or more BAIP tutorials showed z-score gains of .154 units for 5<sup>th</sup> grade students and .172 units for 6<sup>th</sup> grade students. For the 956 5<sup>th</sup> grade students who took 15 or more BAIP tutorials, they were .332 z-score units above the state mean as 3<sup>rd</sup> graders in 2007, but were .486 z-score units above the state mean as 5<sup>th</sup> graders in 2009. Similarly for the 1213 6<sup>th</sup> grade students who took 15 or more BAIP tutorials, they were .233 z-score units above the state mean as 4<sup>th</sup> graders in 2007, but were .405 z-score units above the state mean as 6<sup>th</sup> graders in 2009. As a function of the high power in the statistical tests created by the large sample sizes, these differences in the 2007 and 2009 mean z-scores were statistically significant at the .001 level. As the z-score standard deviations are 1.00, the effect sizes are reflected in the differences between the 2009 and 2007 means, i.e., .154 for 5<sup>th</sup> grade students and .172. These effect sizes are in a range demonstrating low to moderate effects.

To address BAIP effectiveness at the lower grade levels (3 and 4), data were examined at the building level using comparisons of performance for grade level cohort groups in 2007 versus in 2009, i.e., how did grade 3 and 4 students receiving BAIP in a

building do in 2009 compared to grade 3 and 4 students in the building in 2007 when BAIP was not available. For 1691 grade 3 students in buildings where there was evidence of reasonable BAIP usage (at least 70% of students taking tutorials and on average more than 10 tutorials or lessons received by students), the 2009 versus 2007 3<sup>rd</sup> grade z-score means did not demonstrate an increase in performance, but rather showed maintenance of achievement across the two year period (mean z-scores of .246 and .223). However, for 2429 grade 4 students in buildings where there was evidence of BAIP usage, the 2009 versus 2007 4th grade z-score means did demonstrate an increase in performance (mean z-scores of .161 and .235). This effect size differences represent small effects and are statistically significant at the .001 level.