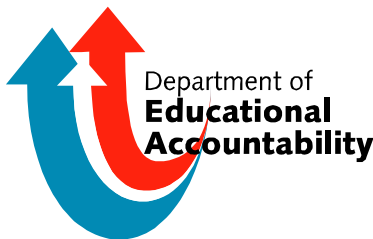


Benchmark Assessment and Reporting Tool (BART) Evaluation

Report I: Alignment, Perceptions, and Academic Growth

July 2004



**Fairfax County Public Schools
Office of Program Evaluation**

BENCHMARK ASSESSMENT AND REPORTING TOOL (BART) EVALUATION REPORT I: ALIGNMENT, PERCEPTIONS, AND ACADEMIC GROWTH

EXECUTIVE SUMMARY

Background

At the beginning of the 2003-2004 school year, Fairfax County Public Schools (FCPS) began piloting two online assessment systems at fourteen schools (two middle schools and twelve elementary schools). FCPS refers to these assessments as the Benchmark Assessment and Reporting Tool (BART) I and II. The BART I system was developed by Tungsten Learning, a division of Edison schools, and the BART II system was created by Princeton Review. BART is to provide schools with a set of electronically administered and scored achievement tests designed to provide accurate and immediate scores. Test results for reading and mathematics in grades 3, 5, and 8 are to be used by teachers to plan instruction, place new students in appropriate courses, and prepare students for the Virginia Standards of Learning (SOL) tests.

The purposes of the BART evaluation are to (a) assist the school division in identifying areas for improving the BART assessments and (b) determine which of these systems is more appropriate for large-scale and long-term use within the division. The evaluation is divided into three reports. Report I (July 2004) investigates the level of alignment (content and performance demands) between each BART system and the SOL, the perceptions of key stakeholders about BART, and the academic growth on the respective BART system. Report II (November 2004) will correlate performance on each BART system with performance on the SOL tests, as well as examine the cost-utility of each BART system. It is anticipated that the school division will use these two reports to decide between the two vendors. Table ES-1 summarizes evaluation results from Report I and shows analyses planned for Report II.

**Table ES-1
Summary of Results Across Evaluation Criteria for Bart I and Bart II**

Evaluation Criteria		Vendor	Rating of Vendor Performance on the Criteria				
			Substantially Below Standard	Below Standard	At Standard	Above Standard	Substantially Above Standard
Report I	Alignment	Tungsten					
		Princeton					
		<i>Scale for rating</i>	Never		Seldom	Most Often	Always/Fully
Report I	Perceptions	Tungsten					
		Princeton					
		<i>Scale for rating</i>	0-34% positive	35-49% positive	50-64% positive	65-84% positive	85% or greater positive
Report I	Academic Growth	Tungsten					
		Princeton					
		<i>Scale for rating</i>	alpha > .05		alpha < .05	alpha < .01	alpha < .001
Report II	Correlation	Tungsten					
		Princeton					
		<i>Scale for rating</i>	0-.59	.60-.69	.70-.74	.75-.84	.85 or greater
Reports I & II	Overall	Tungsten					
		Princeton					

For each evaluation criterion, Table ES-1 shows that each vendor is judged against a pre-determined minimum standard. The standards were communicated to staff and vendors prior to data collection. In addition to comparing each system to the minimum standards, BART I and BART II are compared on the extent to which one system exceeds standards beyond the other, since the goal is to decide between BART I and BART II. Report III (August 2005) will explore how students using the selected BART system subsequently perform on the 2005 SOL tests when compared to matched students who used neither BART system.

Findings

Report I provides findings to vendors and staff about alignment, perceptions, and academic growth. The information may be useful for modifying the implementation or improving the quality and use of assessments prior to the critical summative study represented by Report II.

Alignment

A team of FCPS teachers and curriculum specialists were trained to rate alignment between the BART assessments and the SOLs. The same team was used to rate both assessments. According to this team, Tungsten Learning's and Princeton Review's assessments both met the established criteria for alignment. However, the Tungsten system was perceived as better aligned than Princeton's system at most grade levels across both mathematics and reading. The team also rated Tungsten Learning more favorably on the process it used to develop the FCPS assessments.

Perceptions and Level of Use

Both assessment systems met the minimum standard for overall level of positive reactions from users. However, less than 50 percent of teachers using the Princeton Review system indicated use of assessment data to revise curriculum and instruction and to gauge instructional levels. Moreover, less than 50 percent of those same teachers indicated that the training provided to interpret and utilize the BART data was "good" or "adequate." These ratings were offset by the more positive ratings given by principals, cluster staff, and central office staffs. Based on the student and teacher perceptions, it appears that the Tungsten Learning assessments provide more useful data to inform instruction and help students than the Princeton Review assessments.

Academic Growth

Both BART I and BART II assessments are intended for use as measures of academic growth. If students do not show statistically significant growth from one administration to the next, it is an indication that the assessment may have technical weaknesses (e.g., misalignment of content or item difficulty). Or, the absence of academic growth may indicate a problem with the use of the assessment data. Teachers and others could choose not to use even quality data or use it inappropriately. Either condition could impact inferences about students' academic growth. Therefore, to expose specific concerns with these conditions, an analysis of growth was conducted using BART I and BART II student scores from October 2003 and March 2004.

Repeated measures analyses ($\alpha < .05$) determined whether statistically significant growth had occurred between the two time periods. Analyses indicated that the total student population and all subgroups (as defined by No Child Left Behind) showed statistically significant growth at all

grade levels in mathematics and reading for the Tungsten Learning assessments. This was not the case for the students using the Princeton Review assessments, specifically for the reading assessments. Not only did some groups fail to show progress overall, but on the grade 5 reading October to March assessment all student groups regressed in performance. Similarly, most student groups failed to show growth on the Princeton Review grade 8 mathematics assessment.

Implications of Findings

Report I findings yield implications for preliminary judgments about the Tungsten Learning and Princeton Review assessments. Alignment appears to be a greater challenge for Princeton Review. Without a strong alignment between the assessments and the SOL, instruction can be misguided and subsequent student performance on the SOLs may be unrelated to practice on the Princeton Review assessments. More alignment work is needed for the Princeton Review tests.

Perception and BART usage data suggest that most teachers who are supposed to use the BART I and II assessments are doing so. But, while approximately 59 percent of the Princeton Review teachers and students reported using the data, a greater percentage (80 percent) reported using the Tungsten assessment data. It is possible that time requirements (scheduling students and analyzing data), concern for loss of instructional time, and the systems' downtime due to technical issues all contributed to less than ideal implementation and utilization of the BART assessment data. In the second year, clearer expectations for how to manage these issues and how data are to be used should be communicated and monitored, while considering the barriers to data use that teachers articulated.

The inconsistency of growth from fall to spring indicates problems that require further study, particularly for Princeton Review. The fall and spring assessments are supposed to be at the same difficulty level. However, on the Princeton Review assessment, students performed better in the fall (prior to instruction) than spring which may indicate inappropriate item difficulty level. Princeton Review may need to revise items and assessments to ensure consistency of the level of difficulty of assessments, particularly the grade 5 reading assessments.

Formative Recommendations

Based on the data contained in the full report, the Office of Program Evaluation offers the following recommendations for program improvement and other formative purposes:

1. Vendors should revisit the content of their staff development to ensure its quality and utility with a focus on the utilization of the BART assessment data to inform instruction.
2. Vendors should also develop a staff development component to build central and cluster office staffs' capacity to help schools integrate BART data into instruction.
3. Vendors and Instructional Services should collaborate to review item difficulty to ensure that all assessments meet division expectations.
4. Instructional Services should communicate to principals and teachers expectations for using and monitoring BART data, after considering the barriers to data use described in this report.