District Standards Support Review (DSSR)

A Summary Report

2006-2007

District Goal: Academic Excellence

Ranjana Damle, Ph.D.
Research, Development & Accountability
November 2007

ALBUQUERQUE PUBLIC SCHOOLS

BOARD OF EDUCATION

PAULA MAES
President

DOLORES A.GRIEGO
Vice President

BERNA V. FACIO
Secretary

MARY LEE MARTIN
Policy Chair

GORDON ROWE
District Relations Chair

ROBERT D. LUCERO
Finance/Audit Chair

MARTIN R. ESQUIVEL
Capital Outlay Chair

ELIZABETH EVERITT
Superintendent

LINDA SINK
Associate Superintendent

EDDIE SOTO
Associate Superintendent

RAQUEL REEDY
Associate Superintendent

THOMAS SAVAGE
Deputy Superintendent

RESEARCH, DEVELOPMENT AND ACCOUNTABILITY

930-A Oak Street SE

Albuquerque, New Mexico 87106

(505) 848-8710

www.rda.aps.edu

Rose-Ann McKernan

Executive Director

Instructional Accountability

District Standards Support Review (DSSR)

2006-2007

A Summary Report

Albuquerque Public Schools administration adopted standards-based education for all schools in the late 1990s. The District Standards Implementation Work Group (SIWG)[1], a collaborative group of APS personnel, crafted Standards-Based Education Implementation Plan (SBEIP), a self-assessment rubric, and other guidance material for school staff in 2006. The District Standards Support Review (DSSR) examined standards implementation in schools to identify areas of standards implementation that need more focused District support. DSSR teams comprising APS personnel from many APS departments conducted school observations between November and April of school year 2006-2007. While the District had carried out site visits to evaluate the status of standards implementation in previous years, a new planning team expanded the scope and improved evaluation methods for the 2006-2007 evaluation. Research, Development and Accountability (RDA), along with the DSSR team, created a comprehensive evaluation plan. A collaborative team incorporating Teaching and Learning Systems (TLS), RDA, and other District staff developed standards-aligned observation rubrics and the logistics of carrying out the observations. Conducted in randomly sampled classrooms in each school,[2] DSSR provided the most comprehensive review of standards implementation in APS schools to date.

Methods

The new team brought clarity and focus to what was initially dubbed as 'site visits.' The site visits assumed a new name, 'District Standards Support Review (DSSR),' to reflect the newly defined purpose behind the school visits. The new site visits planning team reconceptualized the purpose of the visits as follows:

- To identify areas in which each school and cluster needs District support to fully implement a standards-based education.
- To create baseline information about standards implementation to measure growth in future years.

RDA and TLS collaborated to develop a formative evaluation plan that designed:

- observation tools and developed observer training
- data collection, management, and analysis procedures
- sampling of classrooms and other logistics of conducting observations, and

[1] The District Standards implementation Work Group (SIWG) is a broadly represented stakeholder group of District and school personnel that worked on the District standards-based education framework.
[2] To minimize disruption to instruction in classrooms, four schools were excluded since they had been observed by district accreditation teams.

- school and cluster reports to reflect the strengths and challenges in standards implementation observed in each school and cluster

Observation Instruments
A subcommittee that included RDA and TLS staff and other District experts in standards-based education created measurements of standards implementation. The team developed observation rubrics that measured standards implementation in three component areas: Instruction, Assessment, and Active Learning. Another component, Communication, was embedded within the other three components. The three instruments, one for each component, were designed to focus direct observation of instruction, review of artifacts, and teacher and/or student interviews in the classroom. The three instruments were indices consisting of between 8 and 12 items, each incorporating a number of indicators of a component of standards implementation. The instruments went through several revisions and included input from experts in the district. Due to a time constraint, the instruments could not be piloted ahead of time or otherwise tested for reliability. However, the evaluation design included the inter-rater reliability testing which will be discussed below.

Sampling
RDA evaluators used random sampling for selecting classrooms although the sample size for each school was smaller than dictated by probability sampling. RDA developed a guideline for sample size using a modified square-root method for each school.[3] Thus, a school with 40 classrooms would require a sample size of six, the nearest square-root.[4] However, this procedure was repeated three times for three components. Thus, in a school with 40 viable classrooms, a total of 18 classrooms were observed – six for each component.[5]

The selection of classrooms for observation was driven by two goals: maximize the number of classrooms within resource constraints and minimize bias by conducting observations in randomly selected classrooms. To achieve these goals, RDA created a list of teachers and a sampling frame for each school. SPSS, a statistical software program, randomly selected a specified number of classrooms for each school. The list provided for randomly selected alternates in case the given teacher was no longer in that school, not teaching one of the content areas, or absent the day of school observation.

Observer Training
Each cluster developed a pool of observers drawing on the cluster teams and instructional coaches. Three training modules for the three components were developed. TLS produced classroom videos for use in training observers for on-site DSSR visits. During the observer training sessions, potential observers watched the classroom videos and

[3] Square-root method of sample selection is taken from Susan Leddick who borrowed it from Edwards Demming.
[4] For logistical reasons, the sample size was rounded to the nearest multiple of 3 (3, 6, 9, and 12). Therefore, depending on the number of classrooms in the schools, the DSSR teams observed 9, 18, 27, or 36 classrooms.
[5] Only the classrooms teaching content areas of math, reading and language arts, and science were included in the count of classrooms.

practiced the use of rubrics to record data. All clusters were trained in a few weeks. Typically, observers were trained to be expert observers in one component.

Inter-Rater Reliability Study Observers
RDA created a pool of observers from RDA, TLS and other administrative departments to function as inter-rater reliability study observers.[6] After a significant recruiting and scheduling effort, inter-rater reliability study observers were trained and scheduled for observation in several schools.

Information Gathering Procedures
The observation procedure was designed for three classroom observations for three periods per observer. Typically, all observations were completed by noon, after three or four class periods. However, one cluster engaged more personnel for each observation and shortened the observation time span even further. The DSSR observations took place in 127 APS schools, in over 1675 classrooms.

Quantitative Observations. All observers used copies of the rubric for each classroom they observed and circled their ratings and wrote comments. To simplify information gathering and to expedite the analysis later, RDA developed scannable forms to record classroom observation. This was supposed to eliminate data entry errors and delays. While there were some unanticipated problems,[7] scan sheets simplified the data gathering and minimized error.

Qualitative Observations. All observers, except for the inter-rater reliability study observers, regrouped at the school after the last observation and discussed their impressions and observations, synthesizing them on a Synthesis Sheet. The Synthesis Sheet listed observed strengths and challenges in the school. These qualitative observations were expected to be informative and of immediate value. Hence, the school administrators received a copy of the Synthesis by the end of the day of DSSR visit.

Limitations of the Evaluation
Due to the time constraint, there was no opportunity to pilot the instruments or otherwise conduct a study of their reliability. Face validity was established through expert review. The observers received training in two half-days, but there was no time for a follow-up review with the observers or debriefing after their first observations to strengthen reliability. The sampling procedure turned out to be not as clean as initially thought since the teacher lists or the sampling frames were incomplete. Finally, RDA and other members of the DSSR team had no control over the quality of actual observations or consistency in practices across clusters.

[6] An inter-rater reliability study tests the reliability of an instrument in which two similarly trained observers make observation using the same instrument. The degree of agreement between their observations is the measure of reliability of the instrument.
[7] Contrary to our assumption, the hand-written information was not readable to the scanner. This resulted in data entry delays, errors, and need for data cleaning.

Data Analysis and Reporting Procedures
The cluster staff gathered all record-keeping materials – scan sheets and rubrics from multiple observers and the Synthesis Sheets – and deposited them at RDA. The sheets went to Computer Services for scanning. The data file was examined for unscanned information and omissions were filled manually. After the data cleaning, individual school reports were generated by the Technology team at RDA. The reports involved average rating scores on each item in each component rubric for individual school. The reports also incorporated the Synthesis Sheet. When all school observations ended, additional reports were generated by cluster and by statistical peer group. [8]

Findings

The school reports were based on the three component rubrics – Instruction, Assessment, and Active Learning – with ratings on each individual item. The school reports provided an average of scores for each item in all three component areas for all classrooms observed. Each school received one of four ratings for each item in the three instruments: Not Evident, Beginning Steps, Nearing Proficiency, and Proficient. After the conclusion of all school visits and subsequent data analysis, RDA completed school reports and reports by cluster and statistical peer group. RDA also produced a report offering schools' modal (most frequent) rating for each component area by cluster and by statistical peer group.[9]

The basic data analysis was designed to discern patterns and derive information useful for the District administration to support standards implementation in schools. Following are significant findings that emerged after the analysis.

Standards Implementation

1. There is substantial variation in standards implementation across APS schools. Most schools have some elements of standards-based education in place.
2. The schools' performance levels for each of the three components are determined by averaging scores on items within each rubric. The analysis of performance levels shows that most schools cluster into two categories - Beginning Steps or Nearing Proficiency (Appendix 1). However there are some remarkable differences across the three components.
    - 60% of the schools are in the category Nearing Proficiency in Active Learning.
    - Schools are almost equally divided in Beginning Steps (46.5%) and Nearing Proficiency (48.0%) in Assessment.
    - 69% of the schools are at the Beginning Steps in Instruction. Standards-based instruction seems to be weak in schools and in need of systematic support.

[8] DSSR school, cluster, and statistical peer group reports are available on RDA/APS website at: http://www.rda.aps.edu/rda/Documents/Publications/07_08/DSSR_0708_Reports.pdf.
[9] See Notes on p. 8 for further details about accessing DSSR reports on the RDA website.

3. The Synthesis portion of the individual school reports identifies the areas of strength and highlights challenges where the school and the District need to focus their efforts, e.g. through continued professional development and other measures. [10]

Process Use–Learning Occurring in the Process of Evaluation
While interacting with cluster staff involved in DSSR observations, two important observations kept surfacing that must be noted.

1. First, the DSSR observers became more aware of and tuned into the nuances of their cluster schools' standards implementation. They understood the strengths as well as areas of growth where they needed to place their efforts.
2. Second, many staff and administrators noted that the observation tools, or the rubrics, detailed for them what proficiency looks like in different areas of standards implementation. In other words, valuable learning about standards implementation occurred during the processes of planning, instrument development, training, and observations.

Standards Implementation and Cluster and Statistical Peer Groupings

1. There is a small difference in standards implementation across clusters. Analysis of variance (compares group means) shows that means of two component scores (Active Learning and Instruction) differ slightly across clusters, although the difference is quite small. This may imply that some clusters are doing a better job of implementing standards in terms of Instruction and Active Learning than others.
2. Analysis of variance shows that means of component scores show little variation across statistical peer groups. Statistical peer groups combine schools with similar demographics that belong to different clusters with their distinctive support systems. Therefore, statistical peer groups comprise significant variation in school means within groups, producing group averages that are similar across the statistical peer groups.
3. Assessment scores do not seem to vary much across clusters or statistical peer groups. This may partly be a result of RDA assessment team's three-year work in unrolling the standards-based progress report across all clusters and schools with extensive teacher training in standards-based assessment.

Component Score Averages and Achievement
The component score means for schools are calculated by averaging scores for all items in each component rubric. These component average scores for schools are then correlated with three achievement scores for 2005-2006 – percent math proficient, percent reading proficient, and Adequate Yearly Progress (AYP) goal met (yes/no).

1. The rating scores for all three components – Instruction, Active Learning, and Assessment – show no statistical correlation to math or reading proficiency.
2. There is a statistically significant but weak positive correlation between Assessment scores and the schools' AYP status. In other words, the schools

[10] DSSR school reports incorporate the Synthesis section of the DSSR team's observations on the 4th page of each school report.

that met AYP in 2005-2006 school year have slightly higher Assessment scores compared to those that did not meet AYP.

Results for Reliability and Inter-Rater Reliability

1. The separate reliability tests conducted on each of the three instruments reveal that each instrument is internally consistent. In other words, within each instrument, items are measuring the same construct and strongly related to each other systematically, and not by chance. The closer is the Cronbach's Alpha value to 1.0 the more reliable the instrument. (See Appendix 2)
    - Instruction – Cronbach's Alpha .916
    - Assessment – Cronbach's Alpha .936
    - Active Learning – Cronbach's Alpha .886

2. Inter-rater reliability is gauged by comparing DSSR observers' classroom ratings with those of the inter-rater reliability study observers'. This study finds that inter-rater reliability is modest for all three instruments. The Assessment Instrument reports stronger correlation coefficients and a relatively greater inter-rater reliability compared to Instruction and Active Learning. (See Appendix 3)
    - Instruction – Eleven out of the 12 items reveal statistically significant correlation values between .54 and .74. One item shows a low correlation of .20.
    - Active Learning – Four out of eight items exhibit statistically significant values ranging between .49 and .59. The remaining items display correlations below .45.
    - Assessment – Seven of eight items present statistically significant correlation values between .62 and .80, with one item at .52.

(Note: The closer the correlation coefficient is to1.0, the better the inter-rater reliability.)

Discussion

The standards implementation scores are weakly related to cluster membership. This finding suggests that some clusters might be providing more robust support for standards implementation than others.

At the present level of standards implementation and the maturity of the DSSR process, the evaluation has not found a statistically significant correlation between standards implementation and math/reading proficiency scores. A few possible explanations are:

1. Low performing schools are getting significant support for standards implementation which is reflected in better ratings on standards implementation.
2. High performing schools, on the other hand, may or may not be implementing standards rigorously.

3. The observation instruments may be measuring some aspects of standards implementation but not others, especially ones that may need longer and deeper observations. In other words, the instruments measure what they are supposed to measure but do not fully capture standards implementation.

Our efforts to support low performing schools in their standards reform may have overshadowed correlations between standards implementation and performance. These findings also imply that other factors may be associated with school performance. Variables such as leadership, mentorship, professional development, and others may produce successful schools in addition to standards implementation. Standards implementation may not be the only path to high performance for schools and hence standards reform must be coupled with these other factors to improve performance.

Recommendations

The DSSR findings may be used to identify the areas of standards implementation that need a boost in terms of professional development and mentorship efforts. Cluster teams may find some common elements in their schools on which to focus. The qualitative data reported in the Synthesis Sheets may be used to get specifics about the schools' needs for support in implementing standards-based education. These findings may be useful for the District administrators in clarifying the proficiency standards and expectations and setting the course for the next school year.

A strong focus on reading and math as the medium of standards implementation coupled with improved measures of implementation could generate a correlation between implementation of standards and achievement in the future.

The DSSR data will serve as a baseline and may be used to appraise future growth in standards implementation. Some of the processes in the evaluation may be refined. The three instruments designed as observation tools will need to be improved for better reliability and validity. The personnel training will have to be restructured. The DSSR process is deeply involving, highly labor intensive, and a strain on District's limited resources. It is recommended that this process be scaled down. For example, only a third of the schools should be sampled in a given year. This evaluation may be conducted every second or third year, with interim years devoted to act on the previous evaluation's recommendations for improvements in standards implementation. The DSSR process may serve APS better with a more focused and in-depth evaluation of a small number of elements of standards implementation.

Notes

All DSSR reports for 2006-2007 are available on the RDA website, under the publication year 2007-2008. These reports can be accessed as follows.

1. District Standards Support Review (DSSR) 2006-2007: A Summary Report

Please find this report at the following web address under the abbreviated title **DSSR Summary Report 2006-2007**:

http://www.rda.aps.edu/rda/Documents/Publications/07_08/DSSR_0708_Summary.pdf

2. DSSR School, Cluster, and Statistical Peer Group Reports 2006-2007

Please find these reports at the following web address under the title **DSSR School, Cluster, & SPG Reports 2006-2007**:

http://www.rda.aps.edu/rda/Documents/Publications/07_08/DSSR_0708_Reports.pdf

These reports can also be accessed by going to the RDA website at **www.rda.aps.edu**, by clicking on "Publications," and going to the heading **"**Published in the 2007-2008 School Year."

Appendix 1

Schools' Performance Levels Based on Mean Scores

Number and Percentage of Schools			
Performance Level*	Instruction	Assessment	Active Learning
Not Evident	7 (5.5%)	4 (3.1%)	1 (0.8%)
Beginning Step	88 (69.3%)	59 (46.5%)	48 (37.8%)
Nearing Proficiency	29 (22.8%)	61 (48.0%)	77 (60.6%)
Proficient	3 (2.4%)	3 (2.4%)	1 (0.8%)

*0.00 - 0.5 = Not Evident; 0.51 - 1.5 = Beginning Steps; 1.51 - 2.5 = Nearing Proficiency; 2.51 - 3.0 = Proficient

Appendix 2

Assessment

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.932	.936	8

Instruction

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.915	.916	12

Active Learning

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.884	.886	8

Appendix 3

Inter-Rater Reliability Table: Assessment

Element	Component	Proficiency Description	Pearson Correlation Coefficient*
Observation	**Assessments**	Assessments evident with student feedback and discussion regarding purpose.	.52 (p<.001)
Artifacts	**Assessments**	Formative and summative assessments which clearly match performance standards and IEP goals were developed prior to instruction and data is used for backward planning.	.71 (p<.001)
	Record Keeping	The evidence of learning (e.g., grade book) is completely organized by student learning outcomes (IEP goals and performance standards) and summative marks are only based on student achievement (progression of student learning toward standard). Assessment is based on most recent evidence of student achievement.	.62 (p<.001)
Student Interviews	**Standards**	Students can describe where s/he is in meeting grade level performance standards and IEP goals, S/he can tell you how they know where they are and why.	.79 (p<.001)
	Multiple Opportunities	Multiple opportunities to learn and assess the standards. There is a high level of student involvement in selection of products and activities. (e.g., DI and MI assessment: Cube or Tic Tac Toe selections and tiered activities).	.58 (p<.001)
Teacher Interviews	**Proficiency**	Student developed rubrics available. Most recent work of the student is used to measure proficiency towards standard.	.80 (p<.001)
	Multiple Opportunities	Multiple opportunities to learn and assess the standards. There is a high level of student involvement in selection of products and activities (e.g., DI and MI assessment: Cube or Tic Tac Toe selections and tiered activities).	.68 (p<.001)
	Standards	The teacher can articulate how the assessment is linked to standards and can articulate how the cognitive level is matched to the performance standard or how it is integrated into instruction.	.70 (p<.001)

*The probability of getting these coefficients by chance is less than 1 in 1,000 or p<.001.

Inter-Rater Reliability Table: Instruction

Element	Component	Proficiency Description	Pearson Correlation Coefficient*
Observation	**Standards**	Standards in student-friendly terms are visible displayed in the classroom (e.g., Big Ideas, Essential Questions, and/or Goal).	.68 (p<.001)
	Standards	Teacher makes students aware of grade level course content standards being addressed and leads class discussions of relevancy of standard being addressed.	.54 (p<.001)
	Standards	More than one strategic learning goal that measures student progress towards performance standards is stated or visually displayed for each core content or class.	.56 (p<.001)
	Instructions	Teacher uses tow or more differentiated instruction strategies specifically based on student needs (e.g., Instruction is differentiated by readiness, interest, or learning profile in content, process or product. Students work in flexible groups as appropriate).	.50 (p<.001)
	Exemplars	All performance levels exemplars are visible and clearly defined (e.g., posting of anchor papers for all levels 1-4)	.56 (p<.001)
Artifacts	**Lesson Plans**	Lesson plans integrate performance standards in tow content areas (e.g., literacy across the content areas, etc.).	0.59 (p<.001)
	Curriculum Maps	Curriculum Maps are developed and aligned to standards They include all five components and are integrated with one or more other content areas (I.e., Performance Standards, Content/Big Ideas, Skills, Essential Questions, and Assessments).	.63 (p<.001)
	Parent Communication	The teacher provides information about standards-based education, and what students need to know and be able to do is communicated weekly to parents/guardians in parent friendly language. Grade level has a clear plan.	.69 (p<.001)
Student Interviews	**Standards**	Students describe more than one strategic learning goal and/or point to a visual display for each core content area/subject goal (e.g., folder, board, wall, or syllabus, etc.).	.74 (p<.001)
	Progress	Students describe or point to all exemplars for all performance levels (1-4).	.20 (p<.176)
Teacher Interviews	**Standards**	Teacher reports that s/he fully aligns all instruction to standards (e.g., backward planning strategies, curriculum maps, essential questions, formative assessments).	.56 (p<.001)
	Collaboration	Teacher reports on-going collaboration to develop, revise and implement curriculum maps to guide instruction throughout the school year (e.g., consensus mapping).	.57 (p<.001)

*The probability of getting most of these coefficients by chance is less than 1 in 1,000 or p<.001.

Inter-Rater Reliability Table: Active Learning

Element	Component	Proficiency Description	Pearson Correlation Coefficient*
Observation	**Expectation**	Teacher involves students in the development of learning goals that they are expected to know and be able to do.	0.59 (p<.001)
	Cognitive Level	Grade level performance standards are referenced by teacher and more than once by students via student dialogue and direct observation.	0.54 (p<.001)
	Cognitive Level	Teacher uses "Why" or "How" questioning strategies to promote critical thinking and elicit students discussions. Teacher uses waiting time effectively.	.49 (p<.001)
Artifacts	**Expectation**	Interactive materials are visible (e.g., KWL charts, categorization, brainstorming or webbing, PDSA, Venn diagrams, T charts etc.).	.43 (p<.001)
	Rubrics	Student and teacher jointly developed rubrics are available.	.51 (p<.001)
	Progress	Students develop with guidance of teacher their own progress monitoring charts and maintain them.	.45 (p<.001)
Student Interviews	**Standards**	Student can explain what s/he is learning and can make connections (e.g., real life applications, prior learning, standards, etc.).	0.32 (p<.023)
	Knowledge of Progress	Student reports how s/he is doing and can provide evidence of their progress and takes ownership of their learning progress by maintaining records.	.41 (p<.004)

*The probability of getting most of these coefficients by chance is less than 1 in 1,000 or p<.001.