

# Improving the Evaluation Method of Abbott Schools

New Jersey Department of Education

Division of Abbott Implementation

FINAL REPORT

October 14, 2005

Review Panelists:

Ronald Ferguson, Harvard University

Andrew Porter, Vanderbilt University

Cecilia Rouse, Princeton University

Staff:

Jun Choi, New Jersey Department of Education

Peter Noehrenberg, New Jersey Department of Education

## **Executive Summary**

This report represents the work of three nationally respected scholars convened by Commissioner of Education, William L. Librera, to conduct an independent review of evaluation work employed by the New Jersey Department of Education (DOE), Division of Abbott Implementation. The review was initiated by the Abbott Evaluation Work Group under the terms set by Abbott X Mediation Agreement by the New Jersey Supreme Court.

The panel recommended a three-part framework to improve the evaluation method of Abbott schools. First, the primary quantitative method, a two stage statistical analysis called a predictive model, can be improved by using multiple statistical models and comparison groups. The integrity of the dataset used could be improved as there is no dependable single metric for student achievement over time and the predictive model is left with cross-sectional data over time which must be re-centered each year. In addition, adoption of a student-level, longitudinal database would dramatically improve the quality of the dataset. Other specific recommendations were made to improve the quantitative method.

Second, randomized controlled trials (RCTs) should be explored to identify practices, that if used by Abbott schools, might improve student achievement, though RCTs would not contribute to evaluating the impact of Abbott investments. RCTs should be adopted cautiously, however, requiring a well-defined, targeted and replicable study of a promising practice. Strong leadership and experience are also required as significant investments in time and money would be necessary to carefully design and implement such a study. If adopted, New Jersey would be a leading state in adopting this innovative research design for education.

Lastly, as research capacity is limited at DOE, continued emphasis on reviewing academic literature and best practices should be made. The work of the Early Literacy, Middle Grades Literacy and Mathematics Task Forces should be continued to bring credible scholarship and practice together to set policy for Abbott schools.

## Overview

The New Jersey Department of Education (DOE), Division of Abbott Implementation, convened a panel of nationally respected scholars from July through October, 2005 to conduct an independent review of evaluation work currently employed by DOE to measure student performance in Abbott schools. The primary work evaluated is a statistical model to compare the performance of Abbott students and schools when compared to students and schools with similar demographic and economic backgrounds. The panel also reviewed qualitative work related to literacy practices in low performing Abbott schools and Abbott schools designated by No Child Left Behind (NCLB) as in need of improvement. Approximately 30 literacy reviews and 105 NCLB reviews have been completed in Abbott elementary and middle schools. Finally, the panel made recommendations to DOE to improve the overall evaluation model of Abbott schools. The evaluation model is focused on whether Abbott schools, given significant increase in resources since the far-reaching New Jersey Supreme Court decisions in *Abbott v. Burke* (Abbott V, 1998), are making a meaningful difference in student achievement and determining what practices are consistently effective.

The independent review was recommended by the Abbott Evaluation Work Group organized under the terms set by Abbott X (2003) Mediation Agreement by the New Jersey Supreme Court. The purpose of the panel is to permit an independent review of the evaluation work going on in the Abbott division by scholars whose reputation is such, that the panel's findings would be accepted as serious, reliable, and meritorious. Scholars were selected based on their experience with large scale datasets and leading qualitative research.

The names of the panel members were recommended by individual members of the Abbott Evaluation Work Group who are themselves scholars recognized for the quality and relevance of their work. While "reputation" is a difficult standard to meet, it is the case with this panel that the DOE sought scholars of the highest national reputation.

Based on this standard, the panelists selected were:

Professor Ronald Ferguson is an economist and Senior Research Associate at the Weiner Center for Social Policy at the Kennedy School of Government at Harvard University. He is also the Faculty Chair and Director of the Achievement Gap Initiative at Harvard University. He is recognized for the depth and richness of his analysis of the achievement gap in education, with particular attention to the differences in performance by African-American and Latino students. His work is frequently cited in publications such as the National Research Council, the Brookings Institution and the US Department of Education. Dr. Ferguson earned his Ph.D. in Economics from MIT.

Professor Andrew Porter is the Patricia & Rodes Hart Professor of Educational Leadership and Policy at Vanderbilt University. He has published widely on psychometrics, student assessment, education indicators and research on teaching. Currently, he has research support from the National Science Foundation and ED's Institute for Education Sciences (Consortium for Policy Research in Education). He is the past-President of the American Educational Research Association and earned his Ph.D. in Educational Psychology at the University of Wisconsin.

Professor Cecilia Rouse is Professor of Economics and Public Affairs at Princeton University and Director of both the Industrial Relations Section and the Education Research Section. Her primary research is in labor economics with a particular focus on the economics of education. She has studied the Milwaukee Parental Choice Program, examined the effects of education inputs on student achievement and Florida's school accountability system. She earned her Ph.D. in Economics from Harvard University and serves on the Abbott Evaluation Work Group.

The panel recommended DOE adopt the following evaluation framework. Their review and recommendations of the statistical model and qualitative evaluation work are discussed within this framework.

**1) Quantitative Review using Multiple Statistical Models and Multiple Comparison Groups**

- 2) **Randomized Controlled Trial in Targeted Program or Practice Areas**
- 3) **Systematic Review of Academic Literature and Best Practices**

### **Quantitative Review using Multiple Statistical Models and Multiple Comparison Groups**

“We don’t know the truth,” said Professor Cecilia Rouse. “The various statistical models are different estimates for getting at the truth.” This is a useful description of the limitations of statistical models employed by DOE and education researchers generally given the lack of high quality data that would allow student-level performance to be monitored over time and the changing nature of statewide assessment programs.

The panel reviewed a two stage statistical analysis called a predictive model used and refined by DOE since 2003. The predictive model assesses the extent to which schools and districts are performing better than expected, about as expected or worse than expected given the composition of their student body based on the state assessment results from 1999 to 2005. The predictive model is based on a regression analysis that generates estimates of subgroup cluster averages after controlling for the effects of school and district membership along with student demographics. Subgroup clusters were defined by student characteristics -- special education, limited English proficient, free or reduced rate lunch eligibility, race/ethnicity and gender. Each student, in turn, was classified as significantly above, significantly below, or indeterminate as compared to a 99 percent confidence interval around their subgroup cluster mean, and a school profile was generated which tabulated the number of students by classification category. In the second stage of the statistical analysis, this profile is compared with all other districts in the state to see if there is a significant difference using a categorical data approach.

The panel noted the predictive model is only as useful as the quality of data that feeds into it. Several challenges stem from the integrity of the statewide assessment data for this type of regression analyses. First, while the baseline year is 1999, the statewide assessment for language arts was restandardized in 2001 and revised again in 2003 when there was a change in vendor. The 2001 restandardization resulted in a 20 point increase

in statewide proficiency rates. The 2003 revision was designed not to change the metric of the test and parallel forms were constructed and equating was done. However, there is not a dependable single metric for student achievement over time, and the predictive model is left with cross-sectional data over time which must be re-centered each year. The statewide assessment needs to be more consistent for the statistical model to be highly dependable.

In terms of data integrity, the panel strongly recommended the implementation of a student-level, longitudinal database so that value-added analyses over time can be performed rather than student-level, cross-sectional data over time. The identification of individual students over time and consistency in measuring their performance is critical to having a control group that results in more reliable statistical measures. Currently, different cohorts of students, whose mobility in/out of the state and across the state, can skew the reliability of the assessment results over time. The ability to follow the same students over time would dramatically improve the quality of the dataset.

The panel found the predictive model satisfactory in terms of its validity and reliability as compared to a generalized linear model, a more standard analysis for education, though some were still uncomfortable with its non-standard approach. The non-standard nature of the predictive model meant that it was not fully tested in a wide variety of circumstances. The panel recommended that multiple statistical tests be adopted to verify the validity of the predictive model. However, the predictive model does appear to be more conservative as the two-stage approach is less sensitive to outliers in the data. In particular, the categorization approach of the predictive model using a Jonckheere-Terpstra (JT) analysis is more conservative than the linear trend of the Least-Squares Dummy Variable (LSDV) analysis.

In addition to using multiple statistical tests to complement the predictive model, the panel made specific suggestions to improve the quantitative techniques used. They recommended that multiple comparison groups be used. A single comparison group, or state mean as used in the predictive model, will always be limited. Other comparison groups could include non-Abbott schools only (as the predictive model includes other

Abbott schools in their state mean calculation), non-Abbott Abbotts (specifically, those schools not designated as Abbott districts but face similar conditions and challenges), and a regression discontinuity approach where comparison schools are selected just above any criteria point. Multiple comparison groups will also shed more light in terms of answering the more fundamental question of whether Abbott schools, given significant increases in resources since 1998, are making a meaningful difference in student achievement.

Lastly, the panel agreed to the following recommendations for improving the predictive model:

1. Drop the Generalized Estimating Equations (GEE) in favor of a Least-Squares Dummy Variable (LSDV) model for generating the predicted values for categorizing students' performance.
2. Drop the District Factor Score (DFS) "community wealth" variable since it is not a dependable and accurate descriptive of schools.
3. Report results from the LSDV regression, and use it to model the effects of interventions and events, while checking the evidence against the outcomes from the categorical data analysis.

In addition to reviewing the predictive model, the panel also made recommendations on improving the overall quantitative methods to improve the evaluation model of Abbott schools. The evaluation framework is focused on whether Abbott schools, given significant increase in resources, are generating a meaningful increase in student achievement and determining what practices are consistently effective. Some analysis has been performed at DOE that compares Abbott districts and non-Abbott districts to shed more light into answering this question.

Figures 1 and 2 contain column charts showing the mean scaled scores for the language arts and mathematics sections of the 4<sup>th</sup> grade state assessments by major subgroups (general education, special education, limited English proficient and total students). The distorting effects of the restandardization of the language arts section in

2001 is very evident as well as the one year decline in Abbott performance after a change in test vendor in 2003, especially among limited English proficient students. Although there is some indication of improved performance, the observed trend of Abbott districts over the period 1999-2005 does not appear to be substantially different from the trend among non-Abbott districts. However, in Figure 2, the improvements in mathematics are clearer for the Abbott districts. The Abbott districts have shown sustained, but gradual improvements over time. There were no changes in the test or instances of rescaling to complicate interpretation of the math trends.

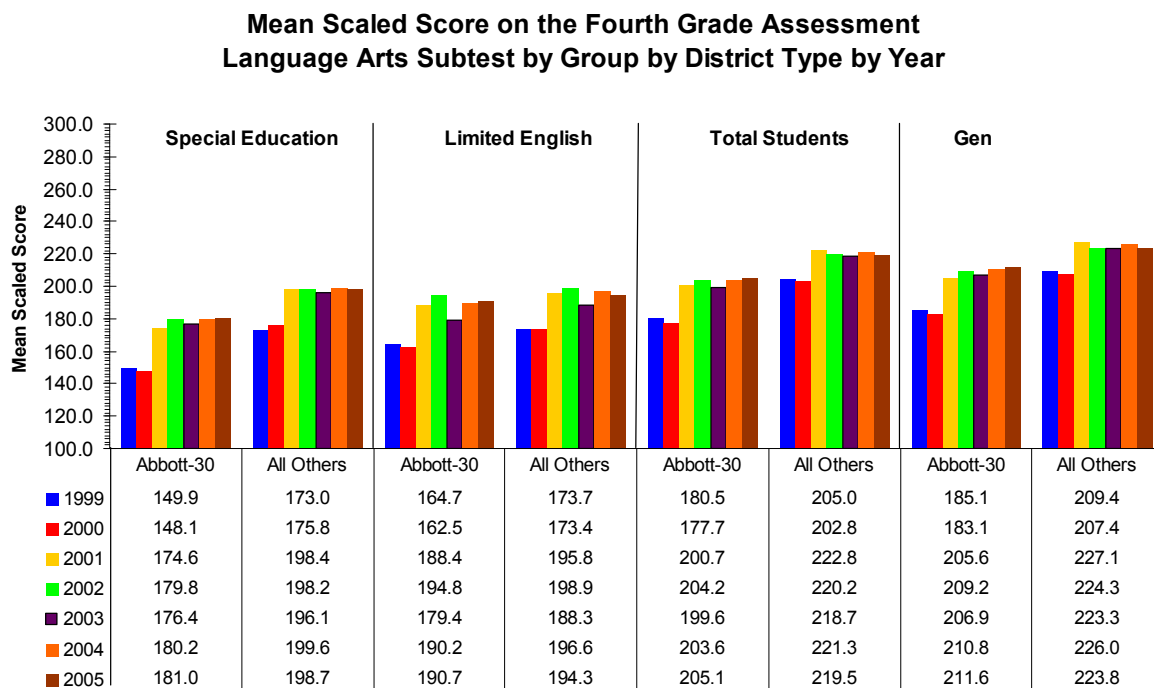


Figure 1

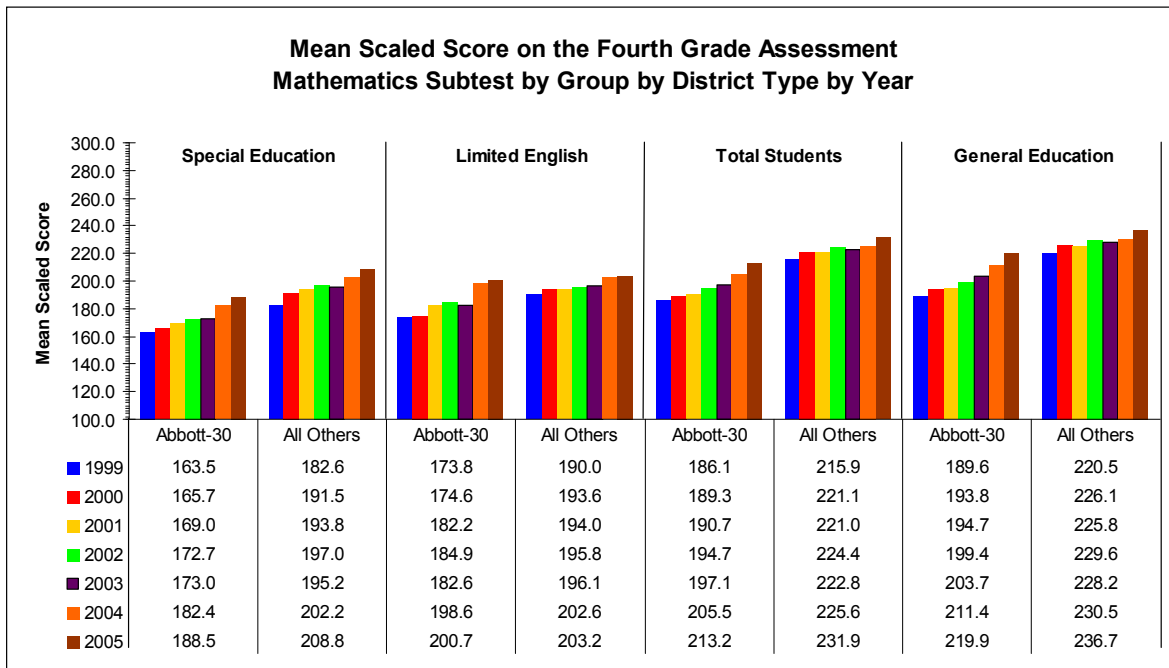


Figure 2

Other recommendations include running more regression analyses to extract the difference between Abbott schools and non-Abbott students after controlling for student demographics and school characteristics without trying to specify the effect of specific interventions. The goal is to identify any overall “Abbott effect.”

In general, however, the integrity of the dataset that DOE uses is limited for the purpose of sophisticated student-level analysis of performance. Professor Andrew Porter, in particular, pointed out the limitations and recommended that DOE not devote extensive time and resources, in addition to the current effort, to analyze these datasets that may not yield much greater meaningful information.

Finally, the panel also reviewed qualitative work related to literacy practices in low performing Abbott schools and Abbott schools designated by No Child Left Behind (NCLB) as in need of improvement. Approximately 30 literacy reviews and 105 NCLB reviews have been completed in Abbott elementary and middle schools. The panel

recommended that the review checklist include not only the “level of implementation,” but also the quality of implementation. For example, under professional development, greater detail should be provided that measures the quality of the program as well as whether the program is targeted to teachers who need them. The panel recognizes that the effectiveness of this qualitative review, however, is largely dependent on the experience and professional judgment of the practitioners who implemented such a review and who provided advice for program improvements in the course of discussions with low performing Abbott schools.

### **Randomized Controlled Trial (RCT) in Targeted Program or Practice Areas**

Well-designed and implemented randomized controlled trials (RCT), commonly used in diverse fields such as medical and environmental research, and respected among scholars in measuring an intervention’s true effect, is gaining momentum in education. “There’s a quiet revolution in education research as a result of randomized trial studies,” according to Dr. Steve Barnett, member of the Abbott Evaluation Work Group and Director of the National Institute of Early Education Research at Rutgers University.

RCTs are studies that randomly assign individuals, classrooms or schools to an intervention group or to a control group in order to measure the effects of the intervention. The resulting difference in outcomes between the intervention and control groups can be attributed to the intervention with a measurable degree of confidence. RCTs can be time consuming and expensive, however, often requiring 2-5 years and multiple millions of dollars to complete for carefully designed and implemented studies. To be clear, the purpose of the randomized controlled trials is to identify practices that if used by Abbott schools might improve student achievement. The exercise would not contribute to evaluating the impact of the Abbott investment.

Given the numerous questions that still remain in regards to the performance of Abbott districts, there would be major challenges in selecting the targeted program or practice area that would be researched. There are various approaches that could be used

to tackle these issues. First, focus the research questions on the largest program and budget areas that would yield the greatest results. The Abbott Evaluation Work Group could develop a prioritized list of research questions based on this determination. Another approach could be to identify practices that had been particularly effective, but questions remain as to whether they can be easily replicated in other schools. Testing the most promising practices on an experimental basis in a carefully designed RCT could yield more definitive conclusions and guidance to educators. After targeted program or practice areas are determined, the study must be well-defined and feasible.

The quality and experience of the principal investigator and leadership team are critical in undertaking a successful RCT. RCTs can be fragile in nature without the right leadership in the field. RCTs must be carefully implemented and the quality of the implementation must be monitored carefully. In some cases, a significant portion of the overall budget – as high as 25-33% - should be dedicated to studying the quality of the implementation. DOE could establish a Research Consortium composed of experienced researchers who have conducted RCTs and New Jersey-based education schools to help implement such a study. A worthwhile study can still fail, so the leadership team and the quality of the implementation are critical to the success of this approach.

### **Systematic Review of Academic Literature and Best Practices**

Since resources are limited at DOE and the capacity of the organization to oversee and conduct high quality research are constrained, greater emphasis should be placed on reviewing the academic literature and best practices available in a more systematic way. Since 2002, DOE has organized the Early Literacy, Middle Grades Literacy and Mathematics Task Forces to bring credible scholarship and practice together to set policy for Abbott schools.

For example, in the case of the Middle Grades Literacy Task Force Co-Chaired by Penelope Lattimer, Ph.D., Special Assistant to the Commissioner, and Dorothy Strickland, Ph.D., Samuel DeWitt Proctor Professor of Education at Rutgers University,

the members conducted an exhaustive review of the research literature, and detailed both best practices and recommendations for action in *Improving the Quality of Literacy Education in New Jersey's Middle Grades*. The best practices were divided into three categories: (1) those that focus on the behaviors and competencies of students in literacy supportive classrooms (Alvermann, et al., 2002); (2) those found in schools and classrooms where students “beat the odds” (perform better than similar students in comparable schools) (Langer, 2000); and (3) those that characterize the professional lives of teachers in schools that “beat the odds” (Langer 1999). In addition, the recommendations for action centered on implementing effective practices, professional development, pre-service teacher education and certification, and assessment.

The panel applauded these efforts and recommended that DOE continue to build on these successful initiatives to bring scholarship and practice together so that the original goals of Abbott to ensure that every child, regardless of background, receives a “thorough and efficient” education can be realized.

## References

Alvermann, D. E. (2002). Effective literacy instruction for adolescents. *Journal of Literacy Research, 34*, 189-208.

Langer, J. (1999). *Excellence in English in middle and high school: How teachers' professional lives support student achievement*. (CELA Research Report Number 1201). Albany, NY: University at Albany, State University of New York, National Center on English Learning & Achievement.

Langer, J. (2000). *Beating the Odds: Teaching middle and high school students to read and write well*. (CELA Research Report Number 1201). Albany, NY: University at Albany, State University of New York, National Center on English Learning & Achievement.