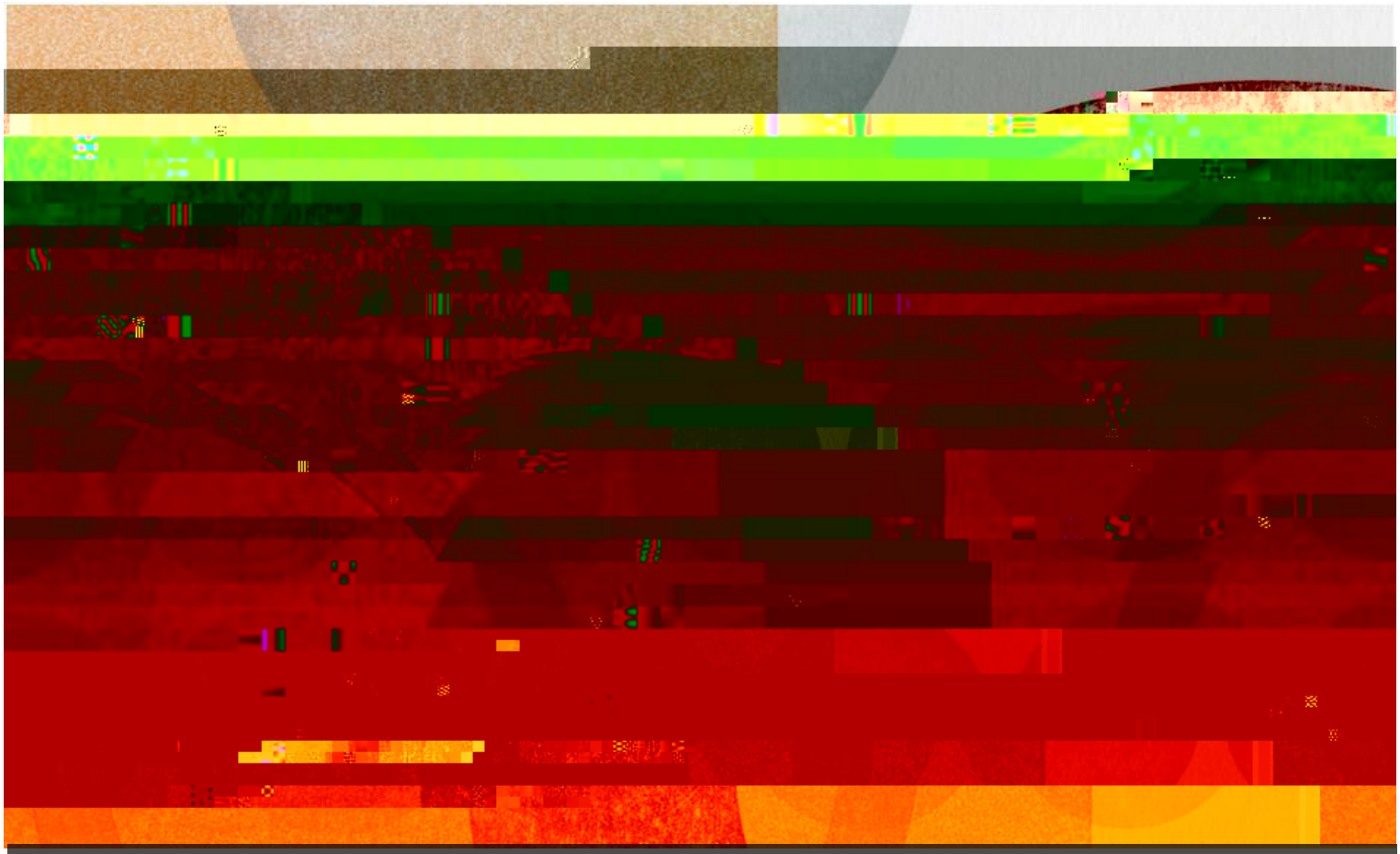

# An Examination of the Effect of a Pilot of Executive Development Program on School Performance Trends in Massachusetts

John A. Nunnery, Ed.D.
Executive Director, The Center for Educational Partnerships
Old Dominion University

Steven M. Ross, Ph.D.
Professor
Johns Hopkins University

Cherng-jyh Yen, Ph.D.
Assistant Professor
Old Dominion University

August, 2010

The Center for Educational Partnerships at Old Dominion University

# The Center for Educational Partnerships

**23529**
**Phone: (757) 683-5449**

## Table of Contents

**List of Tables**



**List of Figures**

matched comparison group, as many of the schools that were selected to participate had been identified as low-performing schools.

The repeated-measures analyses indicated that math effects associated with NISL were consistent across grade levels, so combined effects were estimated. Statistically significant differences in school performance trends were observed on mathematics between NISL and comparison schools and between NISL schools and all schools in the state. In 2006, the students in schools identified for the NISL program scored, on average, -0.17 standard deviation units below students in matched comparison schools in 2006. This difference shrunk to -0.05 standard

*d* estimate of

effect size was +0.10. Results obtained by comparing NISL schools to all schools in

yielded similar results i.e., NISL status was associated with a statistically significant positive trend in aggregate achievement, yielding an effect size estimate of $d = +0.08$.

As with math effects, reading effects were consistent across grade levels so combined effects were estimated. No statistically significant differences in ELA performance trends were observed between NISL and the matched comparison sample. Overall performance on ELA remained virtually unchanged each of the four years in both NISL and comparison schools, yielding an effect size estimate of $d= +0.00$. Likewise, performance on ELA tests in NISL

NISL pilot sample had only one year post-completion prior to 2009 testing, which is probably insufficient time to observe full impact of the program on test scores.

# INTRODUCTION

The establishment of t

Development Program stemmed from the need to assist school leaders in their ability to promote high performance for all students in their schools. The program emphasizes the role of principals as strategic thinkers, instructional leaders, and creators of a just, fair, and caring culture in which

four courses: World-Class Schooling (Principal as a Strategic Thinker and School Designer, Standards-Based Instruction); Teaching and Learning; Developing Capacity and Commitment; and Driving for Results. Designed to be highly interactive, training sessions use simulations and

- applications ( ) to participants.

Prior evaluations of the Executive Development Program prove that the NISL program can be implemented economically and with high fidelity (Meristem Group, 2009). Perhaps more importantly, the research indicates that positive student achievement patterns have been associated with program participation by school leaders. However, these prior studies have used descriptive or correlational designs lacking comparison groups or strong controls over sample selection bias.

More recently, Nunnery, Ross, and Yen (2010) conducted a carefully matched comparison-group ex post facto design to examine NISL program effects in Pennsylvania. Their findings indicate that program participation by school leaders was associated with statistically significant improvement in student achievement for both mathematics and reading over a four-year period. This current study represents a further enhancement in the rigor of the evidence regarding potential effects of the NISL program, as it also is based on an ex post facto, matched comparison design. This interim report provides a preliminary estimate of NISL program effects, with a more sophisticated multi-level modeling of program effects to be completed during the course of 2010.

## METHOD

### *NISL schools*

A total of 131 principals in Massachusetts participated in the NISL program. The analysis sample was restricted to include schools whose principal both completed the NISL program and remained at the same school from 2006 through the end of the 2009 school year. Of the 131 principals who started the program, about 18% ($n$ = 23) did not complete the program due to retirement or transfer, and about 27% ($n$ = 35) changed schools or positions, thus yielding a potential analysis sample of 73 schools. Of these, two were high schools that were eliminated due to the small school-level sample size of NISL schools at this level, and one was a primary school (K-3) which was excluded because test scores were not available for second grade. Finally, five of the remaining schools were not included because complete test score and demographic data were not available for all years for various reasons (being a new school started after 2006, being a school that was closed during the course of the study, etc.). The final sample for the analysis included 65 NISL schools.

### *Dependent measures*

The dependent measures employed in the study were standardized scores (Z-scores) computed from raw scores on the Massachusetts Comprehensive Assessment Program tests in English/Language Arts (ELA) and mathematics. Z-scores were computed separately for each grade level by subtracting the state-mean from each individual student score, then dividing the difference by the state-wide standard deviation. Individual Z-scores were then aggregated across grade levels served by each school, yielding a single school performance index that reflected the mean Z-score for all tested students in each school.

Schools were classified into grade-level types on the basis of the lowest and highest grades served. Schools serving grades three to four, three to five, or three to six were classified as elementary schools. Schools serving grades five-, six-, or seven- to eight were classified as middle schools, and schools serving grades three- or four- to eight were classified as elementary-middle schools.

*Matched comparison school sample*

To facilitate construction of a matched comparison sample, a principal components analysis was performed on the following variables measured in 2006: mean Z-score in English, mean Z-score in Mathematics, the percentage of students receiving free or reduced-price lunch (FRL), the percentage of students receiving special education services (SPED), and the percentage of students with limited English proficiency (LEP). A single component accounted for 63.3% of the variance in these variables across schools. Item loadings were -0.95 for mean English Z-score, -0.94 for mean Math Z-score, 0.83 for FRL, 0.67 for LEP, and 0.50 for SPED. A regression-based factor score was computed from the principal components analysis to yield a composite index to use for matching purposes. Based on the item loadings, higher composite index scores are associated with lower mean achievement and higher rates of FRL, LEP, and SPED.

A matched comparison school sample was constructed by selecting four non-NISL pilot schools with the same grade-level type as each NISL school two with the closest higher composite index scores and two with the closest lower composite scores. For instances in which composite index scores for NISL schools were within this span, comparison schools were selected to yield a total of four comparison schools per NISL school by alternately selecting a comparison school with composite index scores higher, then lower, than those already included.

To assess the adequacy of the comparison school selection process, a 2 (NISL, non-NISL) *X* 3 (Level) *X* 5 (dependent variables) multivariate analysis of variance (MANOVA) was performed with mean 2006 English Z-score, mean 2006 math Z-score, percentage of students receiving free or reduced-

Table 1. Mean Scores on Matching Variables by NISL Status and School Level, 2006

| Level | NISL program completion | English[1] |
|---|---|---|

0.025. Similar procedures were employed for the state-wide comparisons.

# RESULTS

## *Matched-samples Comparison School Results*

### *Math*

Preliminary analyses indicated no NISL status X School Level interaction effects for within-subjects effects ($F_{6,942}$= 0.18, $p$ = .98) or between-subjects effects ($F_{2,314}$ = 0.63, $p$ = .53), indicated a possible violation of the equality of covariance matrices assumption ($F_{10,59173}$ = 2.63, $p$ = .003), so the Greenhouse-

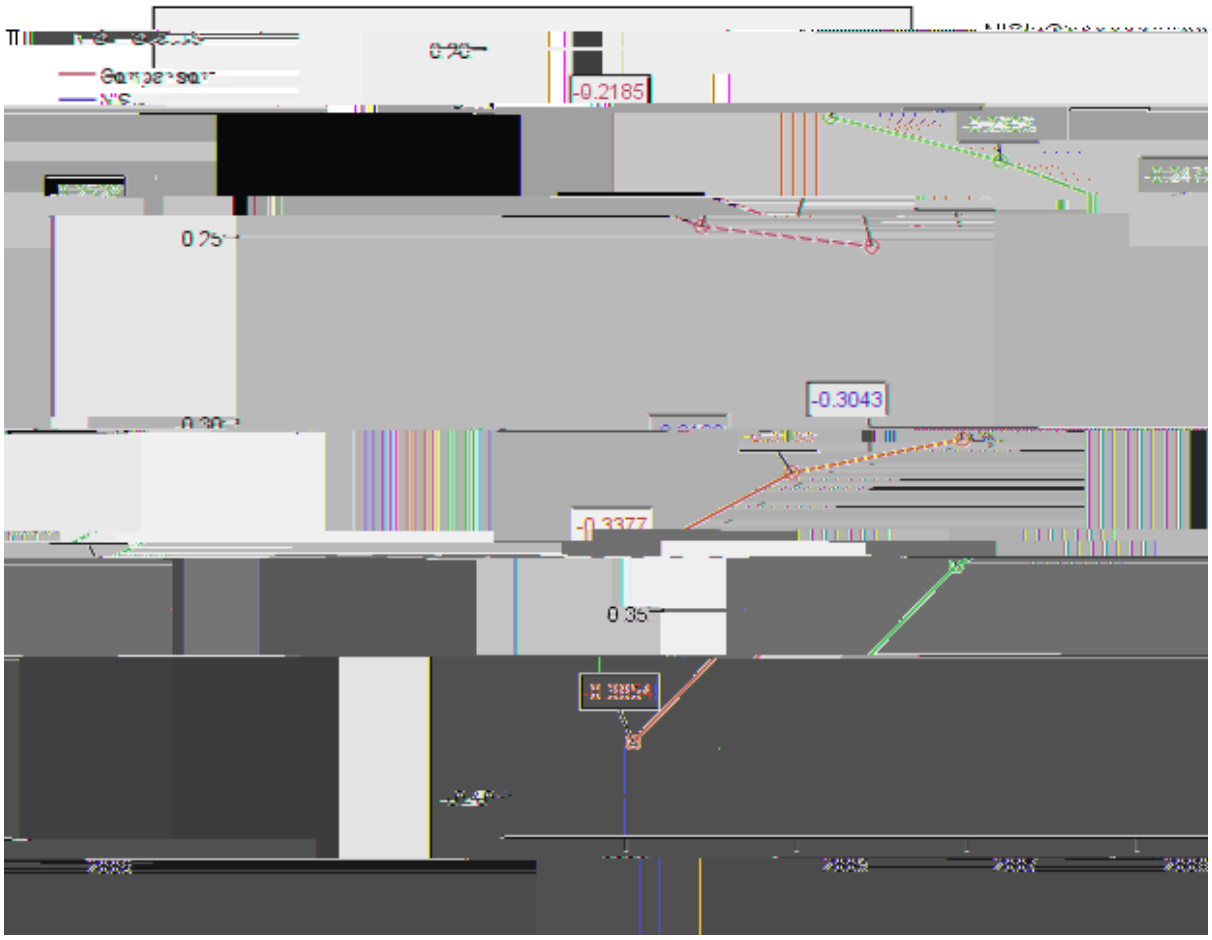


Figure 1. Trends in Mean Math Z-scores in pilot NISL Schools versus Matched Comparison Schools by Year, 2006-2009.

*English/Language Arts*

As with math, preliminary analyses indicated no NISL status X School Level interaction effects for within-subjects effects ($F_{6,942}$= 0.34, $p$ = .91) or between-subjects effects ($F_{2,314}$ = 0.35, $p$ = .70), so analyses were performed using only NISL status as an independent variable.

$F_{10,59173}$ = 1.10, $p$

model were tenable. Tests of the within-subjects interaction between trends in English/Language Arts performance and NISL status showed no statistically significant difference between performance trends in NISL versus comparison schools ($F_{3,954}$ = 0.60, $p$ = .61), yielding an effect size estimate of $d$ = +0.00. Analyses excluding the 12 cohort 2 NISL schools from the

Table 2. Mean Scores by NISL Status and Subject Area with Effect Size

| | 2006 | 2007 | 2008 | 2009 | Effect Size |
|---|---|---|---|---|---|
| Math | | | | | |
| NISL Schools | -.3854 | -.3377 | -.3133 | -.3043 | +.10 (+.12[a]) |
| Comparison Schools | -.2185 | -.2303 | -.2475 | -.2528 | |
| English | | | | | |
| NISL Schools | -.3990 | -.3710 | -.3933 | -.3897 | +.00 (+.01[a]) |
| Comparison Schools | -.2419 | -.2485 | -.2554 | -.2447 | |

Comparison School $n$ = 256; NISL School $n$ = 64. [a]Excluding 12 Cohort 2 NISL schools from comparison sample.

*State-wide Comparisons*

*Math*

Preliminary analyses indicated no NISL status X School Level interaction effects for within-subjects effects ($F_{6,942}$= 0.18, $p$ = .98) or between-subjects effects ($F_{2,314}$ = 0.63, $p$ = .53), so analyses were performed using only NISL status

indicated a possible violation of the equality of covariance matrices assumption ($F_{10,50895}$ = 3.04, $p$ = .001), so the Greenhouse-Geiser correction was performed. A test of the within-subjects effects revealed a statistically significant interaction of trends in mean math Z scores and NISL program status ($F_{2.5,3192.3}$ = 6.91, $p$ < .001). Tests of within-subject contrasts revealed only a statistically significant linear component to the interaction ($F_{1,1277}$ = 12.06, $p$ = .001). As shown in Figure 3, in 2006 students in comparison schools scored an average of +0.38 standard deviation units higher than students in NISL schools, whereas by 2009 that difference was cut to +0.30 standard deviation units. Adjusted 2009 means were -0.04 for comparison schools and

*d* estimate of effect size of *d* = +.08, which was statistically significant at $p < .01$.

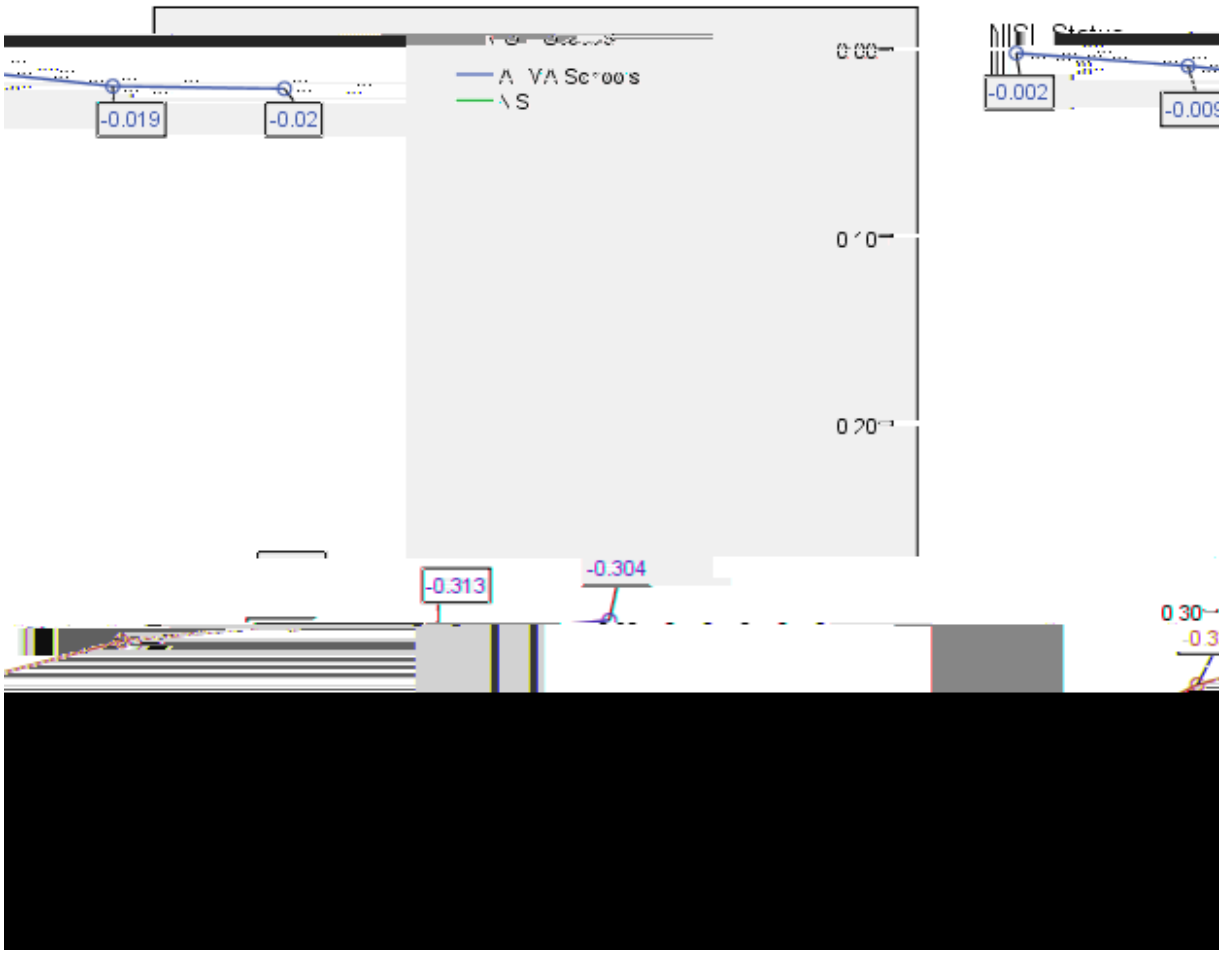


Figure 3. Trends in Mean Math Z-scores in NISL Schools versus All Massachusetts Schools by Year, 2006-2009.

*English/Language Arts*

As with math, preliminary analyses indicated no NISL status X School Level interaction effects for within-subjects effects ($F_{6,3819}$= 0.26, $p$ = .96) or between-subjects effects ($F_{2,1273}$ = 2.39, $p$ = .09), so analyses were performed using only NISL status as an independent variable.

$F_{10,50895}$ = 2.25, *p*

model were tenable. (Note: A conservative

# FINDINGS AND DISCUSSION

## *Study Overview*

This study examined the impact of a pilot of

Raudenbush, 1988; Harris, Kelly, Valentine, & Muhlenbruck, 2000; Rimm-Kaufman, Fan, Chiu, and You, 2007).

In a sense, the lack of a relationship between NISL status and reading trends in the short term lends validity to the observed math effects. The contrast militates against attributing the math effects to selection bias or regression to the mean, both of which potentially are salient threats to the internal validity of the ex post facto design employed in this study. A randomized experiment was not feasible given state and district policies for program implementation. However, the present ex post facto design appears highly rigorous, particularly in minimizing validity threats frequently associated in evaluations of leadership programs with sampling bias. Specifically, participants were described by the state and districts as being mixed in their experiences, success rates, and skills, with some targeted due to demonstrating strong potential for leadership and others due to needing professional development to address weaknesses. Also, the repeated-measures design treated nearly all principals as their own controls in analyzing school achievement patterns over time. Further refinement of the findings, such as examination

## References

Borman, G., Hewes, G., Overman, L., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73,* 125-230.

Bryk, A., & Raudenbush, S. (1988). Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education, 97*(1), 65-108. Retrieved from Education Research Complete database.

Harris, C., Kelly, C., Valentine, J., & Muhlenbruck, L. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monographs of the Society for Research in Child Development, 65*