**Abstract Title Page**
***Not included in page count.***

**Title:**

Making Causal Inferences from a Longitudinal Cluster Randomized Experiment with Crossovers: an Evaluation of a Distributed Leadership Program in Urban Schools

**Author(s):**

Rebecca Maynard, Ph.D., University of Pennsylvania;
Nianbo Dong, Ph.D. Candidate, University of Pennsylvania

## Abstract Body

***Limit 5 pages single spaced.***

**Background/context:**

There is general consensus that school leadership is key to turning around chronically underperforming schools (Herman, Dawson, Dee, Greene, Maynard, & Redding, 2008). However, there is a dearth of empirical evidence pointing to successful strategies for improving leadership among existing school staff. One increasingly popular approach being explored by a number of districts is the distributed leadership model developed by James Spillane (Spillane 2006). This is a strategy for school improvement that is well-grounded in theory, but that has not yet been rigorously evaluated.

An ongoing randomized controlled trial of this school improvement strategy—the Distributed Leadership Training Program (DLT)—was launched in the 2005-06 school year using a complicated lagged, cluster randomized control design. Under this design, schools assigned to the control group in years 1 and 2 were allowed to re-enter the pool for assignment to treatment or control condition in subsequent program years and, moreover, re-entrants into the assignment pool were given higher odds of assignment to the intervention condition than were first-time entrants to the study sample. In order to generate unbiased estimates of the impacts of the DLT program and their standard errors, it is necessary to address both the complexities associated with re-randomizing the control group during the study period as well as the differential odds of assignment to the treatment and control groups across schools and within schools over years.

The second of these issues is addressed by using an inverse-probability-of-treatment weighting (IPTW) (Hong & Randenbush, 2008; Robin, Hernan, & Brumback, 2000). The re-randomization creates a situation that is analogous to the "cross-over" problem (Orr 1998; Bell & Bradley 2008). Because of the randomization to "cross-over" condition in the DLT evaluation, we have an opportunity to test the key underlying assumption of that approach. Bell & Bradley (2008) proposed a non-experimental analytic approach to adjusting for cross-overs for the two-year impact analysis if all the control schools are released to the intervention condition in year 2, which assumes that the first program-year impacts are uniform across calendar years, thereby making it possible to subtract the first-year impact from the second-year outcomes for those being released to the intervention condition to obtain counterfactuals for the two-year impact analysis.

**Purpose/objective/research question/focus of study:**

This study empirically investigates the effectiveness of Distributed Leadership Teacher Training (DLT) program on improving student's academic achievement. In addition, it both tests the assumption that the year 1 impacts are stable across calendar years and examines the importance of properly accounting for the fact that the *standard error* of the first-year impact estimate, which becomes the *measurement error* of the variable (the first-year impact point estimate) used for adjustment in the two-year impact analysis, needs to be taken into account in

the analyses. It discusses the extent to which the findings of this examination of the Bell and Bradley adjustment approach have general applicability.

**Setting:**

The Distributed Leadership Teacher Training Program (DLT) was implemented in a large urban school district. A total of 26 elementary and middle schools participated in the five-year demonstration project.

**Population/Participants/Subjects:**

Any school in the district performing above the 25$^{th}$ percentile on the state assessment used for NCLB reporting was eligible for selection as study sample (required by the Foundation sponsor). Up to 2007-08 school year, there were 26 schools in the study sample, in which 4 were randomly assigned to the treatment condition in year 1 and additional 2 were assigned to the treatment condition in year 2. There are a total of 14,422 students in the study sample, each contributing between 2 and 3 years of data to the analysis.

The average enrollment for schools in the study sample was around 550 students. Roughly two-thirds of the students in these schools are black and 78.2% are economically disadvantaged students. Some schools served students in grades K – 8, while others served only middle-school grades (typically 6 – 8).

**Intervention/Program/Practice:**

The DLT program was designed by a local program team, drawing heavily on the work of Spillane's (2006) and others. It includes instructional leadership training modules on a variety of topics including the following topics: (1) distributed perspective on school leadership, (2) developing professional learning communities, (3) student work and data analysis; (4) collaborative learning culture, and (5) developing evidence-based and shared decision making, etc. In total, leadership teams comprised of between three to six staff from each school, which included the principal and principal-nominated "teacher leaders" received approximately 70 hours of high quality professional development and support a year. In addition, all staff in the program schools received approximately 40 hours of professional development training targeted on needs specifically identified by their school leadership teams. In addition to the formal professional development training, members of the leadership teams received ongoing coaching by dedicated school coaches.

The program theory outlines how the intervention works to finally improve student learning (Figure, 1). Given the contextual factors, the intervention was designed to first improve instruction at teacher leader level (column 3). Second, it is expected that the training and leadership activities will foster collaborative school culture and improve instruction among all the teachers at the school-level (column 4). The resulting changes in school culture and instruction within the whole school are expected to improve student engagement and academic outcomes (column 5 & 6).

**Research Design:**

This study is a longitudinal, clustered randomized experiment. The research design is shown in Figure 2. In 2006-07 school year, 19 eligible elementary schools were recruited for the study sample and four schools were randomly assigned to treatment group receiving the intervention. In 2007-08 school-year, seven new elementary schools were recruited for the study sample. Two out of 22 schools (7 new schools plus the 15 control schools in 2006-07) were randomly assigned to treatment group. Of the 22 schools included in the 2006-07 randomization pool, the 15 schools that had been in the control condition during the 2005-06 school-year were given twice the probabilities of selection as were schools who were newly entering the study sample. The two schools assigned to the treatment condition in 2007-08 were from the 15 control schools in 2005-06.

**Data Collection and Analysis:**

Three years of student demographic and achievement (math and reading) data were obtained from the School District for the 2005-2006 (baseline for cohort 1) through the 2007-2008 (second follow-up year for cohort 1 and first for cohort 2) school years. The student achievement data are results from the state assessments, which are based on a criterion reference test, where each question is matched to content standards and the scores are vertically equated. Cronbach's (1951) $\alpha$ coefficients were found greater than .90 in both reading and mathematics tests.

The *experimental analysis approach*, which is based on randomization, is used for impact analysis in this study. Our estimation proceeds in five steps. First, data of four treatment schools and 15 control schools from the 2005-06 to 2006-07 school year are used to estimate the program impact on student academic achievement ( $M_1^{T1}$) in 2007 (year 1) (notation according to Bell & Bradley, 2008). Second, data of two treatment schools and 20 control schools from the 2006-07 to 2007-08 school year are used for the first-year impact analysis on student achievement ( $M_1^{T2}$) in 2008 (year 2). (Since the original 15 control schools in year 1 were given twice the probabilities of being assigned to the intervention status than the seven newly recruited schools, this estimation uses an inverse-probability-of-treatment weighting (IPTW) (Hong & Randenbush, 2008; Robin, Hernan, & Brumback, 2000).) Third, the null hypothesis $M_1^{T1} = M_1^{T2}$ is tested to see if the first-year impact is uniform across years. This is an empirical test of the assumption in Bell & Bradley (2008). Fourth, the overall one-year impact can be estimated by combining $M_1^{T1}$ with $M_1^{T2}$ within a meta-analysis framework. Fifth, the two-year program impact can be estimated using data of the four original treatment school and 15 control schools from the 2006-07 to 2007-08 school year. Notice that the first-year unbiased impact on student achievement ( $M_1^{T2}$) in 2008 (year 2) must be subtracted from the achievement data in 2008 for the two released schools in order to obtain the pure control group. Furthermore, the measurement error of $M_1^{T2}$ is taken into account in the two-year impact analysis. The estimate of the upper-bound of the 95% confidence interval (CI) of the two-year impact based on the lower-bound of the 95% CI of the first-year impact point estimate ( $M_1^{T2}$) serves as the final upper-bound of the 95% CI of the two-year impact estimate; the estimate of the lower-bound of the 95% CI of the

two-year impact based on the upper-bound of the 95% CI of $M_1^{T2}$ serves as the final lower-bound of the 95% CI of the two-year impact estimate.

A 2-level doubly multivariate repeated-measure approach is used for impact analysis for this longitudinal cluster randomized experiment. We evaluate program impact on two related outcome measures (*MATH* and *READING*). Both measures have three waves (*YEAR*) of data. Students are nested within schools (*j*). The basic form of 2-level doubly multivariate repeated-measure model is given by (notation according to Raudenbush & Bryk, 2002; Singer, 1998):

Level 1 Model (Student):

$$Y_{ijst} = \beta_{0j} + \sum_{k=1}^{2} \beta_{kj}(YEAR_k)_{ijt} + \beta_{3j}(MATH)_{ijt} + \sum_{k=1}^{2} \beta_{(3+k)j}(YEAR_k \times MATH)_{ijt}$$
$$+ \sum_{k=5}^{7} \beta_{(k+1)j}(GRADE_k)_{ij0} + \beta_{9j}(FEMALE)_{ijt} + \sum_{k=1}^{5} \beta_{(k+9)j}(RACE_k)_{ijt} + \varepsilon_{ijst}, \varepsilon_{ijst} \sim N(0, \Sigma)$$

(1)
$$\Sigma = \begin{pmatrix} \sigma^2_M & \\ \sigma_{M*R} & \sigma^2_R \end{pmatrix} \otimes \begin{pmatrix} 1 & & \\ \rho & 1 & \\ \rho^2 & \rho & 1 \end{pmatrix}$$
$$= \begin{pmatrix} \sigma^2_M & & & & & \\ \sigma^2_M\rho & \sigma^2_M & & & & \\ \sigma^2_M\rho^2 & \sigma^2_M\rho & \sigma^2_M & & & \\ \sigma_{M*R} & \sigma_{M*R}\rho & \sigma_{M*R}\rho^2 & \sigma^2_R & & \\ \sigma_{M*R}\rho & \sigma_{M*R} & \sigma_{M*R}\rho & \sigma^2_R\rho & \sigma^2_R & \\ \sigma_{M*R}\rho^2 & \sigma_{M*R}\rho & \sigma_{M*R} & \sigma^2_R\rho^2 & \sigma^2_R\rho & \sigma^2_R \end{pmatrix}$$

Level 2 Model (School):

(2)
$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}(TREAT)_j + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(TREAT)_j + \mu_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}(TREAT)_j + \mu_{2j} \\ \beta_{3j} &= \gamma_{30} + \gamma_{31}(TREAT)_j + \mu_{3j} \\ \beta_{4j} &= \gamma_{40} + \gamma_{41}(TREAT)_j + \mu_{4j} \\ \beta_{5j} &= \gamma_{50} + \gamma_{51}(TREAT)_j + \mu_{5j} \\ \beta_{kj} &= \gamma_{k0}, k = 6,...,14 \end{aligned}$$

$$\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \\ \mu_{2j} \\ \mu_{3j} \\ \mu_{4j} \\ \mu_{5j} \end{pmatrix} \sim N\left[ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & & & \\ \tau_{10} & \tau_{11} & & & & \\ \tau_{20} & \tau_{21} & \tau_{22} & & & \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} & & \\ \tau_{40} & \tau_{41} & \tau_{42} & \tau_{43} & \tau_{44} & \\ \tau_{50} & \tau_{51} & \tau_{52} & \tau_{53} & \tau_{54} & \tau_{55} \end{pmatrix} \right]$$

where,

- $\gamma_{00}$ indicates the baseline test score in *reading* for the control condition
- $\gamma_{01}$ is the difference in *reading* scores at baseline between the treatment condition and the control.
- $\gamma_{10}$ indicates 1-year gain for individuals in *reading* scores in the control condition
- $\gamma_{20}$ indicates 2-year gain for individuals in *reading* scores in the control condition

- The coefficients of the interactions of *TREAT* and $YEAR_k$ (dummy variable) represent the different treatment effects across years, i.e., $\gamma_{11}$ and $\gamma_{21}$ represent the treatment effects in *reading* in year 1 and year 2, respectively, and $\gamma_{11}+\gamma_{41}$ and $\gamma_{21}+\gamma_{51}$ represent the treatment effects in *math* in year 1 and year 2, respectively.
- $\gamma_{30}$ is the estimate of the difference in *math* baseline scores from *reading* baseline scores for the control condition
- $\gamma_{31}$ indicates the additional baseline difference between *math* and *reading* for the treatment condition beyond the control condition.
- $\gamma_{k0}, k = 6,7,8$ are the differences in average test scores in grade 5, 6, and 7 as compared to grade 8 (grades 3 & 4 are excluded, due to the lack of data in previous year(s)). In this model, $GRADE_k$ is a dummy variable for each of four grades on interest in 2008.
- $\gamma_{90}$ is the average difference in *reading* and *math* scores between *female* and *male* (as reference group).
- $\gamma_{k0}, k = 10,...,14$ are the average differences in *reading* and *math* scores for students of different races as compared with white students (reference group)
- $\varepsilon_{ijst}$ is random noise associated with the test in subjects *s* and time *t*. The variance-covariance matrix of this error term captures the association between subjects as well as the association among time (the first-order autogressive [AR(1)] variance-covariance structure is used for temporal dependence in this model, however, other variance-covariance structure is possible, e.g., compound symmetry).

**Findings/Results:**

An early analysis of first year impacts for the first cohort of schools and students showed found some evidence of impacts on teacher-leadership team behavior, but no impacts on student outcomes (Cole, 2008). The reasons that no significant impacts on student outcomes for this early sample could be because the intervention really will not impact student outcomes; but, it also could be because it takes time for the impacts to move from teacher leaders to student outcomes; but it also could be due to the small sample size (not enough power). This second study of the program includes seven more elementary schools and one more wave of data in the impact analysis. The two-year impacts potentially would be larger than the one-year impacts and the larger sample size is has more power to detect the program effects (albeit still not as much as desired). This study will obtain unbiased estimates of one- and two-year impacts of DLT on student's math and reading achievement. Furthermore, Bell & Bradley's (2008) assumption will be empirically examined.

**Conclusions:**

This study will empirically investigate the effectiveness of DLT on improving student academic achievement. Bell & Bradley's (2008) method for handling releasing control groups to the intervention condition will be discussed. Furthermore, this study illustrates an *experimental analytic approach* to analyzing the complex experiment—a longitudinal, cluster randomized experiment with "cross-overs" and unequal probabilities of random assignment of intervention.

# Appendixes

***Not included in page count.***

## Appendix A. References

***References are to be in APA format. (See APA style examples at the end of the document.)***


Bell, S. H. & Bradley, M. C. (2008). Calculating Long-Run Impacts in RCTs That Release the Control Group into the Intervention Condition Prior to the End of Follow-Up. Paper presented at the 2008 SREE Conference. Washington, D. C.

Cole, R. (2008). *The Distributed Leadership Experiment: First Year Impacts on School Culture, Teacher Networks, and Student Achievement.* Unpublished Dissertation. University of Pennsylvania.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Herman, R, Dawson, P, Dee, T, Greene, J., Maynard, R. & Redding, S. (2008). "Turning Around Chronically Low-Performing Schools." IES Practice Guide (NCEE #2008-4020). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved November 29, 2008 from http://ies.ed.gov/ncee/wwc/pdf/practiceguides/Turnaround_pg_04181.pdf

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics. 33(3),* 333-362.

Koger, M. E., Thacker, A. A., & Dickinson, E. R. (2004). *Relationships among the pennsylvania system of school assessment (PSSA) scores, sat scores, and self-reported high school grades for the classes of 2002 and 2003* No. FR-04-26). Alexandria, VA: Human Resources Research Organization.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Robin, J. M., Hernan, M. A. & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology, 11*-5. pp 550-560. Retrieved on October 2nd, 2008 from http://www.jstor.org/stable/3703997.

Orr, L. (1998). *Social Experiments: Evaluating Public Programs With Experimental Methods*. Thousand Oaks, CA: Sage Publications.

Spillane, J. P. (2006). *Distributed Leadership.* San Francisco, CA: Jossey-Bass.

Singer, J. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, 24, 4, 323-355.

## Appendix B. Tables and Figures

***Not included in page count.***

Figure 1. The Logic of the Distributed Leadership Program

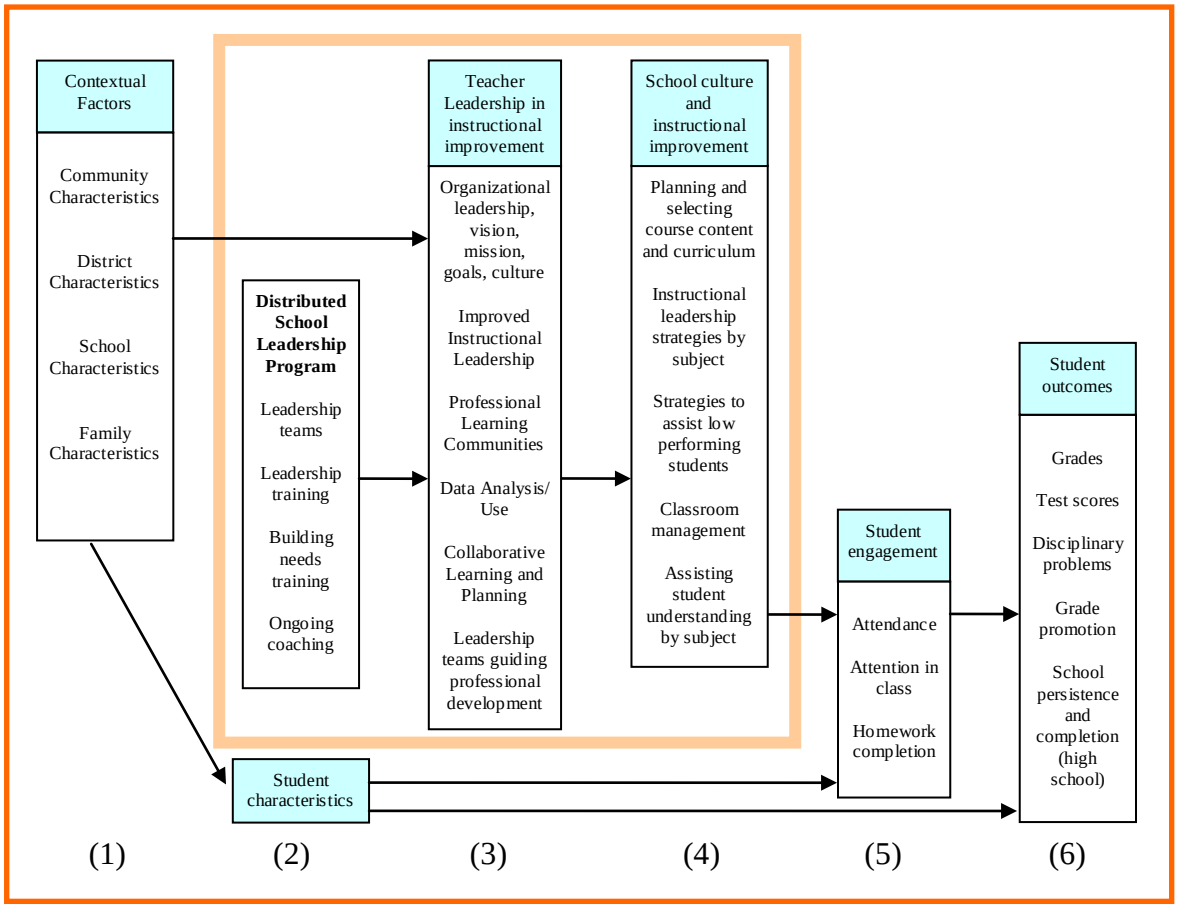

Figure 2. The Research Design of the Distributed Leadership Program