Abstract Title Page
Not included in page count.

Title: Using Mahalanobis Distance Scores for Matched Pairing of Schools in a Randomized Controlled Trial Study of Leadership and Assistance for Science Education Reform (LASER)

Author(s): Dr. Todd Zoblotsky, Dr. Carolyn Ransford-Kaldon, & Dr. Donald M. Morrison, Center for Research in Educational Policy, University of Memphis.

Abstract Body

Background

Developed by the National Science Resources Center (NSRC, a division of the Smithsonian), Leadership and Assistance for Science Education Reform (LASER) is intended to improve the quality of classroom science instruction through a "systemic" approach that engages participants at every level, from classroom teachers up through the highest levels of district, regional, and state leadership. LASER employs the STC science curriculum (also developed by the NSRC), a set of kit-based instructional units that emphasize inquiry-based instruction, i.e., science instruction that engages students in *doing* science rather than just learning *about* science.

LASER has already been the subject of a number of case studies (RMC Research Corporation, 2010; Horizon Research, 2010; Vanosdall et al., 2007). However, experimental studies of the type that might establish a causal link between program implementation, student science learning, and other valued outcomes have yet to be conducted. Also, while previous research has associated inquiry-based science instruction with greater gains in student learning than text-book based methods (Vanosdall, Klentschy, Hedges & Weisbaum, 2007; Banilower, 2007; Ferguson 2009; Bredderman,1983; Shymansky, Hedges, & Woodworth, 1990), only a handful of studies have involved random assignment, and most of these have involved random assignment of students in a relatively small number of classrooms (see Furtak et al. 2009).

With support from the U.S. Department of Education's Investing in Innovation Fund (i3), the Center for Research in Educational Policy has recently embarked on a validation study of the LASER program. The study will involve as many as 27 school districts, including Houston ISD (a large urban district), 7 primarily rural districts in North Carolina, and as many as 19 districts in northern New Mexico, also primarily rural. Eventually the five-year study will involve assessments of developing science literacy in as many as 18,000 students, and classroom observations of 1,800 teachers in more than 120 schools.

Purpose

Data collection is scheduled to occur from Fall 2011 through Spring 2014. The present paper describes the recruitment and site selection process that has been underway since January 2011, with particular emphasis on the use of Mahalanobis distance score to determine matched pairs of sites prior to randomization to treatment and control groups. It is our hope that a reasonably full description of this process will be of interest and practical use to other researchers engaged in similar work.

Setting

The study is taking place in three different parts of the country: Houston, Texas; northern New Mexico, and North Carolina. At this writing, we have just completed the selection process in North Carolina, leading to the identification of 23 Phase 1 (treatment) and 23 Phase 2 (delayed treatment) sites. Assuming that we identify a roughly equal number of schools in the other two regions, we project having a little over 40% of our sites in rural locations, another 20% in towns, 35% in urban areas, and only 4% in suburbs. This is shown in Table 1.

—Table 1 goes about here—

Population

As shown in Table 2, based on data from the identified study sites in North Carolina, and district averages for Houston and the districts in northern New Mexico from which we expect to select our study sites, we project that the students will be 45% Hispanic, 26% Caucasian, 18% African American, 6% Native American, and only 1% Asian. Based the same assumptions, more than 70% of students will be eligible for free or reduced lunch.

—Table 2 goes about here—

Intervention

The National Science Resources Center (NSRC), which is conducting the intervention, has been responsible for site recruitment. At this writing, some 27 districts had been recruited, including Houston ISD, 7 districts in North Carolina, and 19 districts in New Mexico. These districts have all been requested to nominate schools to take part in the study. To become eligible, principals and teachers of science in the nominated schools are asked to complete surveys designed to produce a snapshot profile of the status of science education in the building. As described in more detail below, CREP is using the survey data from the eligible schools, along with other (publically available) data, to create ordered list of matched pairs for each region, which are then randomly assigned to Phase 1 and Phase 2.

Teacher Professional Development and Long-term Support

Teachers in the Phase 1 schools will begin receiving professional development during Summer 2011, and will begin teaching the STC instructional units during the academic year 2011-2012. This will continue through until the Summer of 2014, when the teachers in the Phase 2 schools will begin receiving the materials and training, and formal data collection will cease.

Research Design

The study employs a randomized controlled trial (RCT) complemented by multiple case studies (Baxter & Jack, 2008; Yin, 2003). In the following sections we briefly describe the process we developed for matching pairs of schools in North Carolina.

Data Collection and Analysis

We started with a total of 91 eligible schools in North Carolina, which we defined as nominated schools that had at least a 50% response rate on the teacher surveys. We began the matching process by identifying the school-level variables we felt would be most appropriate. We initially identified a total of 105 variables available to use for matching—primarily publically-available school report card data. We then used a number of strategies to narrow down the list, including using, where available, only the most recent data, and collapsing several groups of individual variables into a single variable. We also looked at correlations between variables that appeared to measure a similar outcome, and where variables were highly correlated, selected the one that made more sense conceptually. In this way, from the initial set of 105 variables available for matching, we narrowed the number of variables down to 18, including three variables taken from our teacher survey designed to measure the extent to which teachers (a) were already engaging students in science inquiry (USE); (b) reported feeling prepared to do so (PREP); and the sum of these two scores (TOTAL).

We next computed the inter-item correlations for our initial 18 matching variables. The inter-item correlations are presented in Table 4. As the three subject area assessments were all highly and statistically significantly correlated (at r=.760 to r=.888), we decided to use only

science proficiency as the matching variable for school-level achievement. We then looked at the remaining variables to identify those with the strongest correlation with science proficiency, and began by limiting inclusion to variables with a statistically significant correlation (either positive or negative) with science proficiency.

Only two teacher-related variables were statistically significant: Percentage of Fully Licensed Teachers in 2009-10 and Teacher Turnover Rate in 2009-10. However, because these two variables were significantly correlated with each other ($r=-.443$, $p<.001$), we used the teacher turnover rate as the one measure of teacher engagement/preparation because the correlation with science proficiency in the school ($r=.332$) was stronger than the percentage of fully licensed teachers ($r=.233$). Finally, although none of the three teacher survey measures were significantly correlated with science proficiency, we wanted to be able to match schools in part on the extent to which teachers were already engaging students in inquiry instruction, or at least felt prepared to do so. As can be seen in Table 5, the three teacher inquiry items were all highly and significantly correlated with each other. Although Use ($r=.201$) had a slightly stronger correlation with science proficiency than Total ($r=.194$), we decided to use Total as the only matching variable from the teacher survey because we wanted to match schools on a more global measure of science inquiry. As a result, we ended up with 11 variables for matching.

In order to pair the schools as closely as possible on the matching variables, we calculated a Mahalanobis Distance score for each school within grade-level bands (i.e., elementary and middle school). The Mahalanobis Distance score, in essence, is a multivariate average score that summarizes a school's distance from all other schools on the mean of all variables included for matching (for definitions and examples see Agodini, Deke, Atkins-Burnett, Harris, & Murphy, 2008; Henderson, Petrosino, Guckenburg, & Hamilton, 2007; Jones, Brown, Hoglund, & Aber, 2010). In other words, all elementary schools were grouped together (i.e., across all seven districts) and all middle schools were grouped together (i.e., across all seven districts).

Final site selections had to take into account two other constraints. First, the NSRC could only serve between 400-500 teachers total (i.e., Phase 1 and Phase 2 combined), so we had to factor in the number of teachers of science at each school. In addition, we wanted to have a representative sample of schools such that each district was represented in the same proportion in the final matched pairs as they were represented in the total pool of 91 schools. In other words, if a district had 15% of the total pool of 91 schools, and represented 12% of the total elementary and 5% of the total middle schools, then that district would be represented in the same proportion in the final matched pairs (i.e., a stratified sample). We also had to ensure that each district had at least one matched pair of schools (which was ensured through the stratified sample). There was also a preference for including scores with lower inquiry scores since the purpose of the intervention was to increase the use of inquiry-based science instruction.

For the initial round of matching, we only included schools with data on all 11 matching variables. This was because schools with missing data did not have a distance score, leaving us unable to determine how similar schools were. For the initial round of matching, we only matched schools within the same district as well. This was necessary to ensure that each district had at least one pair of schools included because some districts were so small that there were not enough schools available to get similar matches, so the best possible match within the district was used.

We used the following order as much as possible to create the matched pairs:

1. Matched schools within the same district with most similar/closest distance scores

2. Matched schools that were the same grade structure (i.e., elementary or middle)
3. Matched schools with most similar/closest distance scores
4. Matched schools with most similar Use/Preparation designation.
    a. We chose a Low Use/Low Preparation designation over other combinations when possible.
    b. We chose a Low Use/Low Preparation designation over having a dissimilar Total Score mean.
5. Matched schools with most similar Total Score mean
6. Matched schools with most similar/closest distance scores
7. Where distance scores were similar, we matched schools with more similar Total Score means first (i.e., Total Score mean took precedence)

A group of three researchers each separately calculated the Mahalanobis Distance scores for all 91 schools. We then met as a group to reach consensus on the schools that should be paired. For each district, we reviewed each of our matches. Where all three researchers agreed, that pair was automatically selected as a match. Where we did not, we each discussed our match and the reason we felt it was the best possible match. In this way, we eventually reached a consensus. Through this process, we selected 23 matched pairs across all 7 districts.

After the final 23 pairs had been established, we placed the 23 pairs of schools into a single list (i.e., a list of 46 schools), then sorted the list of 46 schools by district and their respective pair number (e.g., two schools matched within the same district were given the same pair number). We used the RAND function in Excel to assign each school a random number, then sorted within district and pair by the random number (in ascending order). The first school in the pair was then assigned to Phase I, and the second school was assigned to Phase II.

Results

After making the assignments, we conducted a MANOVA to determine whether there were any statistically significant differences between the Phase I and Phase II schools on any of the 11 matching variables. The MANOVA was statistically significant (Wilks' Lambda=.526, p=.011, Partial Eta Squared=.474). However, on the follow-up univariate analyses, only the percent proficient in science in 2009-10 was statistically significant (p=.004), and the effect size was very small (Partial Eta Squared=.172). Therefore, we feel confident that the two sets of schools are equally matched overall.

Conclusions

Through a systematic winnowing process, we found that we could reduce the number of potential matching variables from more than 100 to just 11 matching variables. By computing Mahalanobis Distance scores for each eligible school, and discussing the results as a team, we were able to identify a set of matched pairs that, when randomly assigned to our two treatment conditions, produced two sets of schools that were well matched on the relevant variables, i.e., represented a level playing field for the RCT in North Carolina. We will use the same process to match and randomly assign schools in our other two regions. We hope that other researchers engaged in similar work will find this account useful.

Appendices

Not included in page count.

Appendix A. References

Agodini, R., Deke, J., Atkins-Burnett, S., Harris, B., & Murphy, R. (2008). Design for evaluation of early elementary school mathematics curricula (MPR Reference No.: 6206-080). Princeton, NJ: Mathematica Policy Research, Inc.. Retrieved April 29, 2011, from http://www.mathematica-mpr.com/publications/pdfs/mathcurriculadsgnrpt.pdf

Banilower, E. R. (2007). Science: it's elementary; year one evaluation report. Prepared for ASSET Inc: Horizon Research Inc.

Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report* , 13(4), 544-559.

Bredderman, T. (1983). Effects of activity-based elementary science on student outcomes: a quantitative synthesis, *Review of Educational Research, 53*(4), 499-518.

Ferguson, G., Long, K., & Kennedy, C. (2009). ASK-IT: assessing science knowledge; implementation through teacher research. Prepared for WA State LASER.

Furtak, E., Seidel, T., Iverson, H., & Briggs, D. 2009. Recent experimental studies of inquiry-based teaching: A meta-analysis and review. Paper presented at the European Association for Research on Learning and Instruction. Amsterdam, Netherlands.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2007). *Measuring how benchmark assessments affect student achievement* (REL 2007–No. 039). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved April 29, 2011, from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2007039.pdf

Horizon Research, Inc. (2010). Evaluation of "Science: It's Elementary." Retrieved from http://www.assetinc.org/documents/ASSETResultsReport2010.pdf

Jones, S. M., Brown, J. L., Hoglund, W. L. G., & Aber, J. L. (2010). A school-randomized clinical trial of an integrated social-emotional learning and literacy intervention: impacts after 1 school year. *Journal of Consulting and Clinical Psychology*, 78(6), 289-842.

RMC Research Corporation (2010). "Washington State LASER: 2008-2009 Evaluation Report," Seattle, WA, RMC Research Corporation. (Submitted to Washington State LASER, January 2010.)

Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance, *Journal of Research on Science Teaching, 27*(2), 127-144.

Vanosdall, R., Klentschy, S., Hedges, L., & Weisbaum, K. S. (2007). "A Randomized Study of the Effects of Scaffolded Guided-Inquiry Instruction on Student Achievement in Science" (Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, Illinois, April 2007.)

Yin, R. K. (2003). *Case study research: Design and methods (3rd ed.).* Thousand Oaks, CA: Sage.

Appendix B. Tables and Figures

Table 1: Projected Number and Percentage of Study Sites by Location

	City		Suburb		Town		Rural		Total	
	N	%	N	%	N	%	N	%	N	%
Houston ISD	46	100%	0	0%	0	0%	0	0%	46	33%
New Mexico	2	4%	5	11%	9	20%	30	65%	46	33%
North Carolina	0	0%	0	0%	17	37%	29	63%	46	33%
Overall	48	35%	5	4%	26	19%	59	43%	138	100%

Table 2: Student Demographics by Location

	African-American	Asian/Pacific	Caucasian	Hispanic	Native American
New Mexico	0%	0%	14%	64%	16%
North Carolina	27%	1%	56%	11%	2%
Houston	27%	3%	7%	62%	0%
Overall	18%	1%	26%	45%	6%

Table 3: Inquiry items from LASER Teacher Survey

How often do students:

1. Conduct science investigations in collaboration with other students
2. Design a science experiment to answer a specific question
3. Participate in field work (e.g., take water samples from local river)
4. Collect data using precise measuring tools (e.g., scales, rulers, thermometers)
5. Write reflections (e.g., in a journal or notebook)
6. Discuss evidence-based explanations in writing
7. Present evidence-based explanations to the rest of the class
8. Discuss evidence-based explanations in small groups
9. Use mathematics to represent and analyze data from a science investigation
10. Conduct exercises in technological design (e.g., robotics)
11. Use computers to collect, represent (e.g., graph), and/or analyze data

How well prepared do you feel to accomplish the following in your science teaching:

12. Use inquiry/investigation-oriented teaching strategies
13. Use eight-week research-based instructional units

Table 4: Inter-Item Correlations for the 18 Initial School Matching Variables

Variable	% Proficient on all Math Assessments 2009-10	% Proficient on all Reading Assessments 2009-10	% Proficient on all Science Assessments 2009-10
% Proficient on all Math Assessments 2009-10	1	.888 **	.760 **
% Proficient on all Reading Assessments 2009-10	.888 **	1	.788 **
% Proficient on all Science Assessments 2009-10 [a]	.760 **	.788 **	1
% Non-White 2010-11 [a]	-.492 **	-.670 **	-.514 **
% Female 2010-11 [a]	.387 **	.321 **	.224 **
% Free & Reduced Lunch 2009-10 [a]	-.551 **	-.726 **	-.538 **
Number of Short and Long-Term Suspensions 2009-10 [a]	-.514 **	-.432 **	-.290 **
Average Daily Attendance Percentage 2009-10 [a]	.563 **	.484 **	.372 **
% Fully Licensed Teachers 2009-10	.457 **	.406 **	.233 *
% Teachers with Advanced Degrees 2009-10	-.047	.032	.136
% with 0 to 3 Years Teaching Experience 2009-10	-.187	-.252 *	-.199
Teacher Turnover Rate 2009-10 [a]	-.467 **	-.524 **	-.332 **
% of Students with a Disability 2009-10 [a]	-.345 **	-.333 **	-.260 *
% of Students who are Limited English Proficiency 2009-10 [a]	-.238 *	-.439 **	-.308 **
Total number of students 2009-10 [a]	.151	.214 *	.208
Teacher Survey: USE Score	.083	.151	.201
Teacher Survey: PREP Score	-.044	-.039	.100
Teacher Survey: TOTAL Score [a]	.059	.122	.194

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

[a] Final variables included in the school level matching process

Table 5: Inter-Item Correlations for the Three Teacher Survey Mean Scores

	Teacher Survey: Use Score Mean	Teacher Survey: Prep Score Mean	Teacher Survey: Total Score Mean
Teacher Survey: Use Score	-		
Teacher Survey: Prep Score	.604 **	-	
Teacher Survey: Total Score	.984 **	.697 **	-

** Correlation is significant at the 0.01 level (2-tailed).