

 **VIGNETTE: Choosing an Evaluator for Rigorous Evaluation**

Purpose: These quotes and materials show how district staff at St. Paul Public Schools (SPPS), MN approached their search for an external evaluator, including how they set selection criteria. Reviewing this vignette and proposal may provide guidance in selecting an external evaluator.

Sources: Interview with Tom Watkins, internal evaluator for SPPS, on August 4, 2008.

Grant application to conduct rigorous evaluation for SPPS, by Geoffrey Borman, PhD, professor at the University of Wisconsin–Madison and independent consultant.

Questions for Reflection

1. What criteria was SPPS looking for in an external evaluator for rigorous evaluation? What additional qualities were necessary for a rigorous evaluator as compared to a regular evaluator?
2. What skills or criteria would your district need to look for in an external evaluator?
3. What process will you use and who will review applications to make the decision on who is best qualified to be your external evaluator?



Choosing an Evaluator for Rigorous Evaluation

Background: Tom Watkins, Internal Evaluator for St. Paul Public Schools, notes some key qualifications the district considered in hiring an external evaluator:

- “Some external evaluators have a much stronger track record of partnership with the district. The evaluator needs to have a strong sense of how we operate and then be willing to adjust accordingly.”
- “We wanted to have somebody who was a high profile partner. Someone who brings a great deal of credibility ... somebody who has published in an area that is either right in this area or near this area.”
- “We also need an experimental design. The design has to define the unique treatment of the schools in our magnet program.”
- “The final report needs to be clear and in a useful format for the magnet staff. It needs to meet all the federal reporting requirements. The evaluator needs to have some credibility with the feds.”

After defining their selection criteria, St. Paul Public Schools district staff solicited applications from several potential external evaluators. Geoffrey Borman, PhD, a professor and independent consultant, submitted this winning proposal.

The Effectiveness of Middle School and High School Science and Technology Magnet Programs in the St. Paul Public Schools

Geoffrey D. Borman
Professor and Independent Consultant
University of Wisconsin–Madison

Policymakers and researchers currently are making unprecedented demands for “scientifically based” evidence on the effects of a variety of educational interventions, policies, and practices (Borman, 2002). True randomized field trials or high-quality quasi-experiments, which compare schools and students receiving a particular intervention to otherwise similar schools and students not receiving the intervention, provide the best evidence for establishing whether or not a program or policy is producing educational effects greater than those achieved by students and schools without the benefit of the program or policy. Applying experimental methods, this analysis will provide the highest-quality, scientifically based research to understand the academic impacts of St. Paul’s Washington Middle School Biosmart science and technology magnet program. Similarly, a high-quality quasi-experiment will inform policymakers and practitioners regarding the potential effectiveness of the Arlington Senior High School science and technology magnet program.

Research Design

To estimate program impacts with the greatest possible precision, we would use a randomized experimental design, which would assign at random students interested in the middle and senior high school magnet programs to a “treatment” or “control group” condition. This type of design is feasible within Washington Middle School. At Washington, the Biosmart science and technology program will operate as a “school- within-a-school” magnet program to which students admitted to the school will be assigned at random. In this case, the counterfactual will be assignment of Washington students to the regular Washington Middle School program. Thus, this experiment will provide important evidence concerning the impacts of the Biosmart program on students’

academic and reported behavioral outcomes. This information will help guide St. Paul policymakers and practitioners with respect to decisions about the future scale-up of Biosmart within Washington and, potentially, to other schools in the district.

In the case of Arlington Senior High School, the conditions are somewhat different. At this time, the district is attempting to generate greater interest in the magnet program. However, the school [Arlington] is clearly not oversubscribed. The lack of strong interest and oversubscription to Arlington makes a lottery for admission, based on random assignment, implausible. Further, Arlington operates a school-wide magnet program and does not have a school-within-a-school intervention—akin to Washington’s Biosmart program—to which students could be randomly assigned. It is not possible to implement a randomized controlled trial within the current context at Washington. Because there are thousands of other similar students who are not being served by the Washington magnet program, though, we may do the next best thing and use a sample of these similar students as quasi-experimental controls (Cook & Campbell, 1979). By identifying similar non-participating students who are matched to the Washington magnet students on variables such as prior achievement, race/ethnicity, free or reduced price lunch status, and neighborhood census tract, we will be in a good position to estimate the value-added effect of the magnet program successfully

Developing an Experimental Assessment of the Educational Effects of the Washington Middle School Magnet Program

The most important threat to the internal validity of any proposed study of the effects of the magnet program is selection bias. That is, the students and families that come forward to participate in the program may differ in important ways from the students and families who do not participate. Such differences, or selection artifacts, make it difficult to know whether the magnet program or the underlying differences among families and students, or some combination of both, really produced the effects of the program that we may find.

The proposed random assignment study will effectively manage the selection bias by randomly allocating students to the Biosmart magnet program or the traditional educational program at Washington. We will begin with nearly a full census of the applicants to the Washington Middle School program.¹ Using random assignment will allow us to draw two samples, a treatment and control group, that are representative of the original population and that are comparable to one another. In assigning students at random, we will form two groups that are, on average, equivalent with respect to things that we can measure, such as baseline achievement levels, motivation, and socioeconomic status, and other things that we may not have measured or even considered. It is not inevitable that the two groups will be comparable, but if the random assignment is done properly and if the samples are large enough, it will result in two groups that will be essentially the same, except that one group receives the intervention and the other does not. From such a sample, we will be in a unique position to generate unbiased estimates of the causal effect of St. Paul’s Washington magnet school programs.

Sample

At Washington Middle School, we will target seventh-grade applicants for the proposed evaluation. The lottery for admission to the Biosmart program, which will be conducted as a randomized process, will provide all students and their families the same fair chance to attend the magnet program. Rather than a “first-come-first-served” strategy, or some other method that favors some individuals or groups, the random assignment process will offer a fair and ethical method for allocating the slots available in the program. There is one clear caveat to mention concerning the randomization process. If there are two or more applicants from one family, both siblings will be subject to the same randomization outcome. In this sense, randomization will actually be conducted by family, with all family members observing the same randomization outcome of admission or denied admission. These procedures related to families will be observed in order to best serve the children and families of St. Paul by avoiding those situations



in which randomization would split up siblings across different programs. That said, though, we anticipate that these situations will be quite rare.

At Washington, the families and students eligible for randomization will be those applying for seventh-grade admittance. Families and students applying through the on-time application process who reside in the district will be eligible for the lottery. In 2004–2005 and 2005–2006, this group consisted of approximately 350 students. With additional future promotion of the Washington program, it is very likely that these applicant numbers will increase.

Data

The primary outcomes measures will be student performance on the district science and math assessments. In addition, we plan to administer student surveys to assess the other strategic goals of the Washington magnet program, including addressing racial/ethnic isolation and improving students' academic engagement in science, and in school more generally. The State of Minnesota and the St. Paul Public Schools administer the following assessments in 7th through 8th grade, which will be the target achievement measures for the evaluation of the Washington program:

- Fall 7th grade Stanford Achievement Test, 10th edition (SAT 10) science and math tests: These assessments will serve as pretest measures of students entering 7th grade science and math achievement;
- Spring 7th grade SAT 10 science and math tests: These assessments will serve as the Year 1 outcomes for students;
- Spring 8th grade state Minnesota Comprehensive Assessments—Series II (MCA-II) science and math tests: These assessments will serve as the Year 2 outcomes for students.

Developing a Quasi-Experimental Assessment of the Educational Effectiveness of the Arlington Senior High School Magnet Program

The Arlington magnet program most directly targets and attempts to attract students and their families. Therefore, our criteria for obtaining the best quasi-experimental matched control group will include important information about students and their families. The criteria we will use to match controls to the Arlington students will be based on the following data from the spring prior to admission to the program:

- Students' 8th grade science scores;
- Student racial/ethnic background;
- Family socioeconomic status, as indicated by free or reduced price lunch data;
- Language spoken within the home;
- Prior middle-school attended;
- Neighborhood census tract information.

Thus, the Arlington applicants and controls will be matched in the sense that they will come from the same middle schools and the same neighborhood contexts. The recent literature on quasi-experimental studies is clear in suggesting that bias is lower when the comparison group is locally matched to treatment or drawn from a control group of a similar or same program at a different site (Glazerman et al., 2002). In addition, the treatment and matched control groups will share similar family backgrounds and achievement backgrounds. Again, the literature on quasi-experimental studies suggests that matching based on pre-treatment measures of the eventual outcomes of the study—in this case, science achievement—reduces bias and improves matching (Glazerman et al., 2002).

Using the district-wide data system from St. Paul, we will be in a good position to identify potential matches for the Arlington applicants. On the basis of baseline student, family background, and school and neighborhood context data for the sample of Arlington applicants and the potential controls, we will perform matching of students using a precise algorithm applied through a

computer-based macro, called “vmatch,” written by Bergstralh, Kosanke, and Jacobsen (1996), following the work of Rosenbaum (1989). The procedure matches treatment cases (in this situation, Arlington students) to control cases to minimize the overall “distance” between the set of treatment cases and the set of control cases. “Distance” in this macro can be defined in a number of ways; we would propose to use the absolute difference in values on the matching variables.

The macro supports both the greedy and the optimal matching algorithms. In the greedy algorithm, each treatment case is matched with a control without replacement. What this means is that after a treatment and control case have been matched to each other, they are removed from further consideration. In contrast, the optimal algorithm will continue to consider the previously paired cases, re-pairing them if it is more efficient to do so. The optimal algorithm is computer-intensive for very large numbers of cases. However, in a situation such as ours, we will be able to perform the optimal matching algorithm efficiently and productively. This method is preferred, in that it improves the matching by 5 to 10% over the results produced by the greedy algorithm (Bergstralh et al., 1996; Rosenbaum, 1989).

We may explore a matching procedure to ensure the best possible matches on one critical criterion, for instance pre-treatment science achievement. In this application, we will match all Arlington students as closely as possible to the control students on the spring of 8th grade science test scores. This method improves matching on this characteristic, but may cause poorer matching on other criteria. A second method we will use will identify the best possible matches on all criteria. Therefore, the first matching procedure, in a way, will weight the science achievement more heavily than the other criteria and the second procedure will weight all criteria equally in the match.

Sample

At Arlington, the analytical sample will include all ninth-grade applicants to the magnet program. Families and students applying through the on-time application process and who reside in the district will be targeted for the evaluation. In 2004–2005, this group consisted of 196 students and included 167 in 2005–2006. As with Washington, we expect that with future promotion of the magnet program, these applicant numbers will clearly increase. In addition, we expect families to apply over the summer and just before the school year begins. These later applicants are approximately the same in number as the on-time applicants.

Data

As with Washington, the key outcomes will be student performance on state and district science and math assessments. Supplemental outcomes will include student survey responses to assess racial/ethnic isolation and academic engagement. The following state and district assessments, administered in 8th through 10th grade, will be the target achievement measures for the evaluation of the Arlington program:

- Spring 8th grade state Minnesota Comprehensive Assessments–Series II (MCA-II) science and math tests: These assessments will serve as pretest measures of students entering 9th grade science and math achievement;
- Spring 9th grade Stanford Achievement Test, 10th edition (SAT 10) science and math tests: These assessments will serve as the Year 1 outcomes for students;
- Spring 10th grade SAT 10 science and math tests: These assessments will serve as the Year 2 outcomes for students;

Final Analytical Sample Size Estimates and Statistical Power for the Washington and Arlington Evaluations

The main analyses proposed for which we need to assess power will be those comparing the achievement outcomes of magnet students and control students. As an absolute minimum, we

expect that a total baseline sample of 300 7th grade students will be randomized to treatment or control for the Washington evaluation. With a minimum of 160 ninth-grade applicants to Arlington and 160 matched controls, the total sample for the senior high school evaluation will be approximately 320 students.

The Washington longitudinal cohort will be tracked from fall of 7th grade through the spring of 8th grade. At Arlington, we will track the longitudinal sample from spring of 8th grade through the spring of 10th grade. Through the use of the central district-wide tracking of students, we anticipate minimal attrition over the course of the study. Only those students leaving the district would have missing information for the study outcomes. Based on prior experience within the district, as a worst-case scenario, one may assume a 85% participation rate and data completion rate at each follow-up. Therefore, assuming that we were to lose approximately 15% of the student samples at the Year 1 and Year 2 follow-ups, the final longitudinal sample for the Washington experiment will consist of approximately 108 students in each of the two conditions, treatment and control, for a total sample of 216 students. For Arlington, the final longitudinal sample will be composed of approximately 230 students, or 115 students in each of the two conditions, treatment and matched control.

The most basic statistical analysis that we will use for assessing experimental effects of the magnet programs is a one-way fixed effects analysis of covariance with 2 levels. The covariate—the spring of eighth grade science pretest for the Arlington evaluation and the fall of 7th grade math and science pretests for the Washington study—will help account for some proportion of random variance and will, thus, help produce greater statistical power to detect experimental treatment effects. To develop an estimate of the proportion of variance explained on the yearly achievement outcomes by the pretests, we examined prior correlations between the various tests administered within the district. Between 7th grade, eighth grade, 9th, and 10th grade, the SAT 10 and MAC II assessments were consistently correlated at approximately $r = 0.75$ to $r = 0.82$. In other words, the pretests explained at least 56% of the variability on the later assessments.

Using this conservative estimate of 56% of the variability explained by pretest and using as a reference the smallest anticipated final sample size estimate of 216 students at Washington, we conducted power analyses for a one-way fixed effects analysis of covariance. The first analysis assumed an effect size (f) of 0.25 and set the criterion for significance (α) at a p -value of .05. This effect size corresponds with that which Cohen (1988) termed a “small” effect, but it is also recognized by researchers as an effect that is of “educational significance” (Slavin, 1990). The analysis of variance was assumed non-directional (i.e., two-tailed). In this case, both the non-adjusted effect of 0.25 and the covariate-adjusted effect would be detected with near certainty.

In a second analysis, which is tabulated below, we computed the power to detect an even smaller effect size of 0.15, which is an effect that is similar to that found in experimental and quasi-experimental studies of class-size reductions (Nye, Hedges, & Konstantopoulos, 1999) and comprehensive school reform programs (Borman et al, 2005; Borman, Hewes, Overman, & Brown, 2003). The power analysis summarized in Table 1 suggests that a non-adjusted effect of 0.15 will be detected with 59% certainty given the assumptions of our design. By using analysis of covariance, though, the expected effect size is increased to an adjusted effect size of 0.23. This covariate-adjusted effect yields quite acceptable power of 0.91. In other words, the ANCOVA analysis, based on a total baseline sample of at least 216 students will yield a statistically significant result with near certainty given an effect of the magnet programs as small as 0.15. Thus, this design should be quite adequate for detecting the expected magnet school effects at both Arlington and Washington.

Table 1. Summary of Statistical Power to Detect Program Effects for Washington and Arlington Magnet Schools



Factor Name	Number of levels	Cases per level	Effect size f	Power	f Adjusted for covariates	Power adjusted for covariates
Magnet assignment	Levels= 2	108	0.15	0.59	0.23	0.91

Within cell SD= 1.00, Variance= 1.00
 Number covariates= 2, R-squared for covariates= 0.56
 Cases per cell= 108, Total N of cases= 216
 Alpha (2-tailed)= 0.05
 Power computations: Non-central F

Qualifications of Principal Investigator

Dr. Geoffrey D. Borman has considerable experience evaluating national policies and programs and using district-level data sets to evaluate the effects of locally implemented educational programs. Trained as a quantitative methodologist at the University of Chicago, Dr. Borman (Ph.D., 1997) is a Professor of Education at the University of Wisconsin–Madison, the Deputy Director of the University of Wisconsin’s Predoctoral Interdisciplinary Research Training Program, a Senior Researcher with the Consortium for Policy Research in Education, and the lead analyst for the Center for Data-Driven Reform in Education at Johns Hopkins University.

Professor Borman’s main substantive research interests revolve around social stratification and the ways in which educational policies and practices can help address and overcome inequality. His primary methodological interests include the synthesis of research evidence, the design of quasi-experimental and experimental studies of educational innovations, and the specification of school-effects models. Over the past seven years, Borman has led or co-directed seven major randomized controlled trials of education interventions. He has conducted three recent research syntheses, including a meta-analysis of the achievement effects of 29 nationally disseminated comprehensive school reform models. Finally, other ongoing projects reveal the consequences of attending high-poverty schools and living in high-poverty neighborhoods and uncover some of the mechanisms through which social-context effects may be manifested.

Professor Borman has been appointed as a methodological expert to advise the National Research Center on the Gifted and Talented, three of the nation’s regional educational laboratories funded by the Institute of Education Sciences, the national Center for Comprehensive School Reform and Improvement, and several other national research and development projects. He is a Principal Standing Panel Member of the Education Systems and Broad Reform Research Review Panel of the U.S. Department of Education, Institute of Education Sciences and was recently named to the 15-member Urban Education Research Task Force established to advise the U.S. Department of Education on issues affecting urban education. Dr. Borman was the recipient of a 2002 National Academy of Education/Spencer Postdoctoral Fellowship Award, the 2004 Raymond Cattell Early Career Award from the American Educational Research Association, and the 2004 American Educational Research Association Review of Research Award.

Proposed Work Plan and Budget

I propose to lead three primary tasks, which are described below: (1) study design; (2) implementation of randomization process; and (3) analysis of outcome data.

Study Design

My work involving the study design will begin with analyses to help determine the appropriate sample sizes needed for the study. These analyses will assess the statistical power and the overall feasibility of alternate study designs. Specifically, this work will develop key parameters of the study that are required for estimating statistical power, including the anticipated effects that



we will observe, the amount of variability in those effects across students and classrooms, and the extent to which future performance on achievement tests can be predicted by past performance. The resulting analyses will help determine how many students are needed for the proposed study. This activity will solidify objectives for recruitment of the families and students to take part in the study.

The first deliverable will be targeted toward the SPPS and the additional stakeholders involved in the magnet program. The deliverable will describe the outcomes of the statistical power analysis, will propose a student recruitment plan, and will outline the overall study design and its implications for all parties involved. I anticipate that this deliverable will be presented to the parties involved and will be prepared as a paper.

Implementation of Randomization Process

Next, I will negotiate a randomization plan with the SPPS team and the local educational and policy stakeholders from the school involved in the study. This plan will be designed in such a way as to accommodate the realities of the magnet program implementation and the needs of the school and families. However, the plan must also be implemented in such a way that the design will produce true random assignment of students to the magnet and control conditions. This process will clearly benefit from my experiences negotiating the implementation of similar random assignment studies with program developers and local practitioners and policy makers from Baltimore to Los Angeles, and many places in between.

The second deliverable will articulate the process of randomization that has been negotiated. It will be developed as a memorandum of understanding that will be agreed upon and entered into by all parties (i.e., researchers and local practitioners/policymakers). The memorandum will help define how the randomization process will take place and be maintained over time and will hold all parties accountable for specific actions that will help maintain the integrity of the randomization process and the causal inferences that can be made from the study's results.

Analysis of Outcome Data

Personnel from the locales involved and/or others appointed by SPPS will manage all aspects of data collection, with consultation provided by me. All students in both the treatment and control conditions will be assessed according to the local and state testing policies. These staff will collect and compile all data. The data will be provided to me in an Excel or SPSS file and will be in a format that is ready to analyze. I also plan to collaborate on some analyses with researchers from the SPPS.

Yearly analyses will document the progress of the study and will track the progress of the treatment and control students. These analyses will provide evidence that bears on the integrity of the randomization process, the integrity of the implementation of the design, the adequacy of the longitudinal follow-up and data completion rates for both treatment and control groups, and will document the magnet program effects during each year, and longitudinally. I will prepare written reports of the effects during each year of the contract. These reports will be submitted to SPPS and will be submitted to the nation's premier educational research journals, such as *Educational Evaluation and Policy Analysis* and the *American Educational Research Journal*. Further, the results of the study will be disseminated through presentations at professional conferences, such as the annual meeting of the American Educational Research Association. Finally, information and papers will be submitted to other publication outlets targeted toward practitioner and policymaker audiences, including *Educational Leadership* and *Education Week*. This dissemination strategy will help publicize the effects the SPPS magnet program to a wide audience of practitioners and policymakers who make choices regarding the implementation of educational programs and products, and to a wide audience of educational researchers and scholars interested in rigorous evidence on school-based interventions.



Level of Commitment and Budget

Geoffrey Borman will commit a total of 30 days per year over the three years of this contract. These 30 days will include time for consulting on the design, randomization plan, analysis of the data, and preparation of final reports. Dr. Borman's consulting rate is \$1,800 per day, and the total yearly rate for Dr. Borman's participation in the project is \$54,000.

References

- Bergstralh, E.J., Kosanke, J.L., & Jacobsen, S. J. (1996). Software for optimal matching in observational studies. *Epidemiology, 7*, 331-332.
- Borman, G.D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education, 77*(4), 7-27.
- Borman, G.D., Hewes, G.M., Overman, L.T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*, 125-230.
- Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005b). The national randomized field trial of Success for All: Second-year outcomes. *American Educational Research Journal, 42*, 673-696.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2d ed.) Hillsdale, NJ: Erlbaum.
- Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Glazerman, S., Levy, D.M., & Myers, D. (2002). *Nonexperimental replications of social experiments: A Systematic Review*. Princeton, NJ: Mathematica Policy Research, Inc.
- Nye, B., Hedges, L.V., & Konstantopoulos, S. (1999). The long-term effects of small classes: A five-year follow-up of the Tennessee class size experiment. *Educational Evaluation and Policy Analysis, 21*, 127-142.
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*, 1024-1032.
- Slavin, R.E. (1990). IBM's Writing to Read: Is it right for reading? *Phi Delta Kappan, 72*(3), 214-216.

ⁱ A select groups of students, including some special education students with severe disabilities, will be excluded from the random assignment process and will be assigned by staff at Washington to the educational programming deemed most appropriate for their educational needs.