eMINTS NATIONAL CENTER

Evaluation Team Policy Brief

November 15, 2004

An Investigation into Selection Bias in FY02 eMINTS Schools

Adam Bickford

This report analyses the impact of the nonrandom selection of eMINTS schools, teachers and students in the FY02 eMINTS cohort. The report summarizes the evaluation-design choices made by the program in light of the simplistic study design endorsed by advocates of "scientifically based research." It also proposes a modeling strategy to estimate the potential impact on MAP scores of the nonrandom selection of teachers and students. The analyses suggest that no identifiable selection biases exist among the schools, teachers and students participating in the FY02 eMINTS cohort.

Introduction

The eMINTS program began in 1999 as an initiative sponsored by the Missouri Department of Elementary and Secondary Education (DESE) and the Missouri Research and Education Network (MOREnet). Its purpose was to demonstrate how ubiquitous classroom access to a multimedia computing environment and the Internet could enhance instructional practice and improve student learning and test performance. The eMINTS evaluation project completed its analysis of the first cohort of eMINTS schools, the FY00 cohort, in 2001. In this cohort, and in each cohort since, analyses of Missouri Assessment Program (MAP) tests have shown that students enrolled in eMINTS classrooms scored significantly higher than students enrolled in other classrooms in the same schools. This apparent success has led DESE to support the installation of eMINTS classrooms throughout the state of Missouri and has led to the creation of the eMINTS National Center in the spring of 2004.

As the eMINTS program established itself and as its first evaluation findings were published, the standards for educational evaluation were being critically revised in conjunction with the passage of the No Child Left Behind Act of 2001. From the perspective of this legislation and the developing standards for "scientifically based research," the MAP score differences presented in the eMINTS evaluation reports are not necessarily evidence of a successful program. According to such standards, the MAP-score differences observed between students enrolled in eMINTS classrooms and students not enrolled in eMINTS classrooms could be the result of any number of factors unrelated to the installation of multimedia classrooms in participating schools and the ongoing support of the eMINTS professional-development program for participating teachers. The fact that the eMINTS program does not randomly select its participating schools from the population of Missouri elementary schools, coupled with the fact that it

This report is one product of the eMINTS Evaluation project. See the following website for other reports and the overall evaluation plan: http://www.emints.org/evaluation.

The eMINTS Evaluation focuses on student impacts, teacher impacts, changes in learning environments and the outcomes of project services.

does not randomly assign teachers or students to eMINTS classrooms, has led to questions about whether the eMINTS experience actually improves student performance. This report investigates whether or not the nonrandom assignment of eMINTS schools, teachers and students had any measurable impact on student scores for the 2003 MAP Communication Arts and Mathematics tests in the schools participating for the FY02 cohort.

This report summarizes the eMINTS evaluation design, describes the outcome measure (the MAP tests) and then discusses the design in light of the issues raised by the call, embedded in the No Child Left Behind legislation, for randomized experimental studies of educational programs. Finally, the last two sections analyze the potential impact of selection bias on eMINTS MAP-score differences, first at the school level and later at the classroom level.

The eMINTS Evaluation Design

The eMINTS evaluation design is a quasi-experimental design that uses a state-sponsored standardized test to gauge program effectiveness. The overall design and its measures were chosen in response to resource constraints and the character of the educational policy environment in Missouri. This section outlines the key features of the design and its measurement strategies in light of those constraints. The design is then placed in the context of "scientifically based research," a set of standards that favor randomized designs in educational studies (see National Research Council, 2002). The discussion of the eMINTS evaluation design in light of issues of randomization demonstrates that the simplistic approach to randomization and selection contained in the standards endorsed by advocates of "scientifically based research" does not account for the complexity of evaluation in actual educational settings.

The eMINTS Evaluation Design

Formally, the design used in the eMINTS evaluation analyses is an example of a "posttest-only design with nonequivalent groups" (see Shadish, Cook and Campbell, 2002: 115-117). While this form of evaluation is not the strongest possible research design, it is the most rigorous evaluation design possible given the operational constraints of the eMINTS program.

Data to complete this design comes from three sources: individual student-level test-score data from the state MAP test archive, individual student data records from participating schools and direct classroom observation of instructional practices. The first source of data, the individual student MAP record, provides the dependent variable (a student's total MAP score) and several classification variables, for example, whether or not a student has an IEP.

The second source of data, student data collected from participating schools, allows for the linkage of MAP records to individual classrooms and the determination of a student's enrollment in an eMINTS or non-eMINTS classroom by identifying a student's teacher

of record. This information is not directly available from the state MAP-test archive, as the archive lists a student's test proctor in a restricted field and the proctor listed may not be a student's classroom teacher. Consequently, this information must be collected directly from participating schools.

The third source of data, direct classroom observation of instructional practices, allows for the classification of instructional practices in terms of the eMINTS instructional model[1]. This instructional model is derived from an understanding of constructivist educational principles and is taught in the eMINTS professional development program, a two-year series of training seminars eMINTS teachers complete as part of their participation in the program.

In every program cohort, the eMINTS evaluation project has worked to collect all the available data from all the participating schools and classrooms. Over the course of the eMINTS program, the eMINTS evaluation project has received data from all participating schools. This reporting provides some assurance that the eMINTS evaluation project's data represents the most complete and accurate picture possible of the participating schools. Furthermore, such reporting allows for a complete accounting of student MAP-test performance in the participating schools.

The Outcome Measure: The MAP Tests

The primary outcome measure used in the eMINTS evaluation is a student's total score on a Missouri Assessment Program (MAP) test, a standards-based assessment administered statewide in grades 3 and 4, grades 7 and 8 and grades 10 and 11. These assessments include constructed-response items and performance events in addition to more conventional multiple-choice items. The MAP tests in the elementary grades are administered in pairs: communication arts and science in the third grade and mathematics and social studies in the fourth grade.

The use of standardized-test scores as outcome measures has many disadvantages (Reckase, 2004). Some of these disadvantages limit the utility of the test for investigating how student enrollment in an eMINTS classroom improves student learning. First, the structure of the tests, which assess different subject matter areas in different years, makes any sort of longitudinal or value-added assessment of performance impossible. Due to this feature of MAP testing procedures, it is not possible to discuss the improvement of student scores at the classroom or student level. While most districts maintain a schedule of off-year non-MAP testing to supplement the MAP, these tests are too diverse, and are administered in too many different ways, to use in a cohort-wide program evaluation.

A further issue in the use of standardized-test scores as outcome measures is the meaning of the outcome score. The MAP test provides only a single, broad measure of student

[1] Classroom observations produce several important qualitative data findings, most notably, the eMINTS Lesson Typology and the eMINTS Classroom-Climate scale. These findings are described in the following reports: *A General Typology of eMINTS Lessons* (2001) and the *eMINTS Classroom Climate Scale* (Tharp, 2003 and Tharp, 2004). These reports are available at the eMINTS-evaluation website: http://www.emints.org/evaluation.

performance. While the test itself is aligned to state standards, it does not allow investigators to examine particular skills. For example, one of the third-grade MAP tests is communication arts, which encompasses reading, reading comprehension and writing skills. Students in eMINTS have scored significantly higher on this test than non-eMINTS students. However, it is not possible to quantitatively determine whether students in eMINTS classrooms read better, comprehend more or write better than their non-eMINTS peers..

An additional problem arises when linking MAP-test performance to classroom instructional practices. The eMINTS program encourages teachers to adopt inquiry-based instructional practices using the available technology. The scheduling of the MAP test, typically once per school year in the early spring, complicates the linkage between observed instructional practices and student outcomes. In practical terms, a single classroom-period observation of instructional practices must represent all of a teacher's instructional practices over an entire school year. Although the evaluation project employs a methodology that helps to produce an empirically verifiable account of classroom instructional practices and seeks to triangulate observational findings through a series of supplemental rubrics and open-ended interviews, the project does not have sufficient resources to fully document the range of instructional practices teachers employ. At its best, the observation of instructional practices undertaken by the eMINTS evaluation project can only be used in the most general sense. As with a MAP test itself, classroom observations can only identify gross differences. They cannot measure which practices are more effective in raising MAP scores than others.

Nevertheless, the MAP tests are the official state assessment and for a project wishing to inform state educational policy it is essential to use these tests in documenting student performance relative to the eMINTS intervention. Since the inception of the eMINTS program in 1999, Missouri's entire district accreditation process has focused on student performance as expressed by MAP scores. For eMINTS to be part of that process, it was necessary to adopt the MAP test as its outcome measure.

Positive and Negative Characteristics of the eMINTS Evaluation Design

The next two sections consider the specific characteristics of the eMINTS evaluation design as it has developed since 1999. The discussion highlights the complexity of the research task and places many of the design decisions in the context of the overall program.

In assessing the overall project design, it is helpful to keep in mind that the eMINTS evaluation project is a program evaluation, not an independent research project. The contrast between a research project; where researchers have effective control over treatments, group characteristics, selection of outcomes, etc.; and an evaluation project; where researchers must work within the constraints of an existing program; is important to consider when judging the adequacy of the design choices made by the evaluation team. The activities of the eMINTS evaluation are a response to decisions and requests made by the eMINTS program. Among these decisions were the decision to focus on aggregate classroom outcomes within individual program cohorts, the decision not to

commit program recourses to develop an independent comparison group in favor of maximizing the number of eMINTS classes in individual cohorts, and the decision to measure outcomes at the end of a teacher's participation in the professional development program. Some of these decisions supported the evaluation effort, while others did not.

Positive Characteristics of the Overall Design

The central concern of the eMINTS evaluation project is the comparison of the MAP scores of students enrolled eMINTS classrooms with the MAP scores of students not enrolled in eMINTS classrooms. This comparison has several positive features. First, analyzing data from all students in a given grade in a participating school controls for school differences.

Secondly, this comparison exploits the naturally occurring classroom structure of a school. Elementary students are enrolled in discrete classes and the instruction that students receive varies among the different classes in a grade. This structure provides for an apparent set of differences. The eMINTS evaluation project exploits this structure by characterizing eMINTS classrooms as a treatment group —the treatment being the different ways of teaching and learning in a multimedia classroom—and using the non-eMINTS classrooms as a naturally occurring control group.

The third positive aspect of this design is its ease of implementation. This design does not require researchers to construct a comparison group of matched classrooms in other, non-eMINTS schools. The cost of constructing such a control group, which would entail doubling the number of schools within a given cohort, looms as a perennial concern for a project that wishes to maximize its resources by funding additional eMINTS schools and classrooms.

Lastly, the relative ease for schools in meeting the data requests from the evaluators works in the project's favor. The evaluation team asked each cohort of eMINTS schools to provide a set of student data from its buildings' student-information systems. Rather than relying on schools to produce student records from particular classrooms, the current design collects information about all students in a given grade (grades three and four). This step simplifies the data-submission process and minimizes the evaluators' interactions with the schools. The best proof of this approach rests in the fact that over the course of the first three eMINTS cohorts, all schools have submitted the requested data.

Shortcomings of the Overall Design

Although the use of such a straightforward design in the eMINTS evaluation project has some particular advantages, it has many disadvantages as well. The most obvious disadvantage is the role of nonrandom selection and the potential for bias inherent in the selection process. Nonrandom selection raises several additional issues related to the measurement of instructional practice and student achievement.

The key selection issue concerns the place of randomization in the design. In considering whether the results of the eMINTS evaluation project have validity, namely whether or

not the observed MAP differences between students enrolled in eMINTS classrooms and students not enrolled in eMINTS classrooms can be attributed to the application of the eMINTS professional-development program and the installation of eMINTS multimedia classrooms, examining the methods by which the different study groups were selected is essential.

Employing a random selection process yields equivalent groups. . In the logic of experimental design, elements in the treatment group experience some intervention, for example, teacher-participation in the eMINTS program, while elements in the control group do not experience the intervention. Assuming that the treatment and control groups are equivalent at the beginning of an investigation, one can attribute any differences between groups at the end of the investigation to the intervention. For example, if an investigator were to randomize student assignment into an eMINTS classroom at a beginning of a study, that investigator would, at the end of the study, be able to attribute any differences observed between those students enrolled in the eMINTS classroom and those students not enrolled in the eMINTS classroom solely to the effects of the eMINTS enrollment, as the two groups of students could have been presumed to be equivalent when the randomized classroom assignment took place.

This very simple characterization of randomization is at the core of the ideas about "scientifically based research." Proponents of this standard of research maintain that any analysis that does not rely on randomization in establishing comparison groups cannot make any valid claims about the efficacy of an intervention because the analyst cannot assert that the groups were equivalent at the beginning of a study (see, for example, the study review standards used by the What Works Clearinghouse, 2002). From this proposition it follows that any observed differences in nonrandomized designs could be due to existing differences between the treatment and control groups and attributing such differences to the intervention would be improper. In addition to this formulation of causation, this research standard implies that the investigator can control the process of selection.

While situations where an investigator can control the assignment of subjects into control and treatment groups may be common in purely academic research settings, they are less common in applied and policy-related settings, such as in the eMINTS evaluation. The eMINTS program and the eMINTS evaluation project have several areas where the model of randomization and the standards of "scientifically based research" are inappropriate as guiding methodologies.

The first area where such methodologies are unsuitable is the selection process itself. Instead of being a simple matter of randomly choosing students for eMINTS classes, the eMINTS program has three levels of selection, each with its own peculiar constraints: school selection, teacher selection and student selection.

School selection

The first level of selection happens at the school level. eMINTS classrooms are resource-intensive rooms, requiring the installation of new equipment (most notably an electronic

whiteboard and high-lumen projector), the provision of a dedicated T100 Internet connection for each classroom, the purchase of classroom computers and, often, the purchase of new classroom furniture to accommodate the computers. The magnitude of this initial expense makes randomly selecting schools to participate in the project unreasonable.

Instead, the program has managed school selection through a competitive application and grant process since its second cohort in 2000. The Missouri Department of Elementary and Secondary Education (DESE) conducts the competition, choosing eligible schools from the population of non-eMINTS schools and inviting them to apply for places in each year's cohort. Most applicants receive assistance from DESE in completing their application. The competition regularly awards thirty to forty grants to new eMINTS schools in each program cohort. This practice is consistent with DESE's stated program goal to establish eMINTS classrooms in at least one school in every Missouri school district. However, these schools cannot be characterized as being selected through a random process.

Once a cohort of schools has been established, one possible opportunity for randomization would be to stagger equipment installation and professional development among selected schools as a means of setting up equivalent treatment and control groups of schools. In theory, one could randomly select a subgroup of schools to begin the program in the next school year, with the remaining schools beginning the program at a later date. However, DESE does not see this as a workable option, preferring to maximize the number of schools that begin the eMINTS program in each cohort.

At the school and district level, participating schools are clearly self-selected. Typically, self-selection introduces any number of biases, as self-selected schools will likely differ from the population of schools statewide. Creating a synthetic control group by sampling non-eMINTS schools through a process known as "propensity score matching" (Rubin, Stuart, and Zanutto, 2004) would, potentially, address this bias. In the first cohort of the project, the eMINTS evaluation team proposed this strategy, progressing to the point of drawing together a matched sample of non-eMINTS elementary schools with which to compare eMINTS schools. The expectation was that the eMINTS evaluation team would apply the same data-collection strategies (observation of instructional practices, collection of student data from the schools, use of MAP scores and so forth) for both groups of schools. This process would have created equivalent control and treatment groups. However, the eMINTS program rejected this strategy as being too expensive.

The issues of school selection and of applying procedures to account for self-selection biases highlight an important difference between the eMINTS program and its evaluation and an independent research effort. The focus of the eMINTS program has always been on maximizing the benefits for Missouri's school children. This focus has led to program policies that support funding the maximum number of schools in each cohort. This priority has led to the rejection of proposals to randomize the selection of schools, whether at the point of initial selection or at the point of the award of eMINTS grants, and to the rejection of design features that would address selection biases at the school

level, such as observing instruction from a matched sample of non-eMINTS schools. When faced with a choice to serve more of Missouri's students or to develop a more academically robust evaluation design, the eMINTS program has always chosen to deploy its resources to maximize the number of students enrolled in eMINTS classrooms. In a program designed to test a research hypothesis, such resource decisions would support a more rigorous evaluation design.

Teacher selection

The second level of selection happens among teachers. The eMINTS program began in the elementary grades, grades chosen for the relative simplicity of instructional delivery. In the elementary grades, the MAP test is administered in grades three and four. Consequently, eMINTS classrooms were originally limited to grades three and four. Programmatically, the number of eMINTS classrooms was also limited to two project-funded classrooms per school, although individual districts could supplement these two classrooms with their own funds. Over the history of the program, participating districts have funded approximately one-third of all eMINTS classrooms.

The eMINTS program expects its teachers to participate in a two-year professional-development program designed to introduce both technology skills and constructivist instructional practices. The bulk of this 200-hour program is scheduled for out-of-contract time, typically in the evenings.

This training requirement means selected teachers must sacrifice their own time not only to learn to use the eMINTS technology in their teaching but to increase their understanding and application of constructivist teaching practices. Consequently, selecting teachers randomly and then expecting them to complete the program presents multiple difficulties. Instead, teachers are nominated by their schools. In the school-application process, DESE requires that nominated teachers submit statements outlining their understanding of the eMINTS program and its professional-development commitment. Like the school-selection process, the teacher-nomination process could introduce a selection bias. In particular, one might assume that a school would select its best teachers to participate in the program.

In small schools with one only section each of grades three and four, no grounds for selection bias exist, since all eligible teachers are selected by default. However, larger schools have grounds for teacher-selection bias. The eMINTS program has always allowed schools to follow internal procedures in selecting teachers, and the process of selection can vary considerably within a cohort of schools. Some schools select teachers based on seniority, other schools conduct lotteries, while some schools sponsor formal application processes within the school or district. In practical terms, schools manage teacher selection in so many ways that the biases introduced by any particular method of selection likely cancel each other out at the cohort level. However, these biases may still have relevance at the school level. Most frequently, given the evidence that students in eMINTS classrooms typically score higher on the MAP tests than other students, the question of whether or not to attribute such successes to the overall ability of the teachers selected for eMINTS classrooms arises.

Student selection

The third level of selection occurs in the placement of students into eMINTS classrooms. In the simplistic model of randomization employed by advocates of "scientifically based research," this level seems like the most obvious area to randomize. Under such a model, one would assign students to eMINTS or non-eMINTS classrooms using some sort of random process. However, such an approach would present at least two problems in an investigation with an intervention that lasts for an entire school year.

The first problem concerns the classroom assignment itself. While the initial randomization of classroom assignments and equivalent design groups at the beginning of a school year may not pose insurmountable difficulties, maintaining this equivalence over the course of the year poses unique challenges. Virtually every elementary school in any of the eMINTS cohorts has had students change classrooms in the course of the year and has had to account for student midyear withdrawals and enrollments. One investigator working with a single school might be able to account for all changes in student populations and still demonstrate some level of design-group equivalence at both the beginning of an investigation and at its end, but this process is extremely difficult to monitor in a cohort of thirty schools.

The second problem deals with school and parental autonomy in classroom assignment. Requiring a school to assign students to classrooms according to a research protocol places the researchers, or the program sponsoring the research, in the position of running the school. Such a protocol would mean ignoring all the local expertise of the teachers and administrators in determining the best classroom assignments for particular students. It would require the school to ignore the wishes of parents in the education of their children. One might assume that all parents would want their children to be enrolled in eMINTS classrooms, however discussions with principals indicate that a substantial minority of parents have asked that their children not be placed in eMINTS classrooms. The reasons for this type of request rest with individual parents, but eMINTS schools would have to reject these requests in order to achieve truly random classroom assignments.

To avoid these problems, the eMINTS program decided to rely on individual schools to manage the student-assignment process. In most cases, schools do not alter their traditional assignment policies for the eMINTS classrooms. Principals in eMINTS schools consult with teachers and parents before making classroom assignments just as their non-eMINTS counterparts do. In isolated cases, schools have implemented policies that prevented students from being enrolled in eMINTS classrooms over two consecutive years. These policies have helped maximize the chances for all students to spend a year in an eMINTS classroom. However, none of these policies help address the possibility that the students in eMINTS classrooms could differ from the students in other classrooms.

The eMINTS Evaluation Design and "Scientifically Based Research"

The eMINTS evaluation project began reporting its first major findings just as states began to implement the No Child Left Behind legislation. Part of this implementation

was the development of the set of research standards used to identify "scientifically based research" outcomes. Many of these standards focused on an experimental paradigm for determining outcome validity. As the previous discussion suggests, these standards do not account for the complex reality of eMINTS schools and should not be used to assess the adequacy of the eMINTS evaluation design.

As a result of the multiple constraints placed on it by the eMINTS program, the eMINTS evaluation design is extremely simple. The constraints introduce selection biases at every level of the investigation. The eMINTS evaluation team works with an awareness of these biases and has attempted to account for the shortcomings of the basic evaluation design where possible. For example, the analyses of the eMINTS evaluation team have shown a consistent relationship between MAP scores and the instructional practices of eMINTS teachers. Generally, the students in eMINTS classrooms where teachers consistently apply the constructivist, student-centered principles embodied in the eMINTS instructional model score higher than students in eMINTS classrooms where teachers apply more traditional, teacher-centered instructional models. This finding has led the eMINTS evaluation team to conduct classroom observations in *all* classrooms in an eMINTS school to help disentangle the impact of the eMINTS equipment and instructional practices in supporting test performance. The first year for such observations was during the FY03 cohort. Analyses of the results from this cohort should help account for teacher-selection biases by specifying and statistically controlling for different approaches to classroom instruction.

The standards of "scientifically based research" offer a challenge to improve the practice of educational evaluation. Many of the activities of the eMINTS evaluation project do not meet these standards. However, the eMINTS evaluation team conducts its work with an understanding of its limitations. As the eMINTS evaluation team understands the constraints under which it operates, the eMINTS evaluation project has made the accommodations it believed were necessary to both inform the program and conduct defensible analyses of standardized-test data, in order to make a report such as the current one possible.

Assessing Selection Bias in the FY02 eMINTS Cohort: School Selection, Teacher Selection and Student Selection

The following discussion offers a comprehensive assessment of the grounds for the nonrandom selection of schools, teachers and students in the FY02 eMINTS cohort. The first analysis describes the school-level performance of the FY02 eMINTS schools relative to a matched sample of non-eMINTS schools. This analysis does not attempt to compare classroom-level differences between the two groups of schools. Rather it compares three years of building-level MAP scores in order to assess the extent to which the FY02 eMINTS schools differed systematically from the sample of non-eMINTS schools.

The second analysis attempts to model teacher and student selection biases directly by estimating a series of multilevel selection models based on the work of James Heckman (1979, see also Winship and Mare, 1992 and Fu, Winship and Mare, 2004). These models show some evidence for selection bias among teachers but suggest that these biases may not differentially impact the performance of eMINTS and non-eMINTS students.

Each of these analyses has its basis in MAP tests in the two core areas specified by the No Child Left Behind legislation: communication arts and mathematics. The analysis of school differences considers the school-level performance of FY02 eMINTS schools and a matched sample of non-eMINTS schools over a three-year period. This period begins with the 2000-2001 school year, when the FY02 schools applied to participate in the eMINTS program, and ends in the 2002-2003 school year, when the FY02 eMINTS teachers completed their professional-development program. The analysis of teacher and student differences focuses on data from the 2003 MAP tests.

The FY02 eMINTS Cohort

The FY02 eMINTS cohort consists of 39 schools with eMINTS classrooms in the third and fourth grades. Of the cohort schools, 23 have eMINTS classrooms in both third and fourth grades, 3 have eMINTS classrooms in the third grade only and 13 have eMINTS classrooms in the fourth grade only. One of the schools is a university-sponsored laboratory school, and 4 schools serve minority populations in Missouri's two main urban centers, Saint Louis and Kansas City.

This analysis uses MAP data from a total of 3416 students in 180 classrooms. Approximately half the students in the analysis were enrolled in eMINTS classrooms. The overall MAP-score difference between students enrolled in eMINTS classrooms and students not enrolled in eMINTS classrooms is 5.94 points on the third-grade MAP Communication Arts test and 8.45 points on the fourth-grade MAP Mathematics test (see Bickford, 2004).

Assessing Selection Bias at the School Level

The assessment of selection bias at the school level involves a retrospective analysis of MAP scores over a three-year period ending in the 2002-2003 school year. The first year of this period, 2000-2001, is the year that the FY02 eMINTS schools applied to the program. The second year, 2001-2002, is the first year of the FY02 eMINTS teachers' participation in the professional-development program. The third year, 2002-2003, marks the end of the FY02 eMINTS teachers' participation in the eMINTS professional-development program.

Assessing whether the FY02 eMINTS schools differ statistically from a sample of non-eMINTS schools requires data aggregated at the school level. This analysis uses a different measure of school performance. The next section describes this measure, the MAP Performance Index, a school-level outcome measure from the Missouri School Improvement Program (MSIP) district-accreditation process.

The MAP Performance Index

The MAP Performance Index is a measure drawn from the percentage distribution of the five summary levels of student performance derived from the total MAP score. This index assesses school and grade-level performance in the Missouri School Improvement Program (MSIP) district-accreditation process. Due to the usefulness of this measure in aggregate performance assessments, the following analysis uses the MAP Performance Index exclusively.

The MAP Performance Index is based on the distribution of the MAP Performance-Level classifications within a school. The MAP Performance Level is a five-category scale ranging from Step 1 for the lowest level through Advanced for the highest level. The MAP Performance Index has a range of two points: a 1 indicates that 100% of all students scored at the Step 1 level, a 3 indicates that all students scored at the Advanced level.

The MAP Performance Index reflects the percentage distribution of MAP Performance Levels at a specific tier of aggregation. For example, the analysis of selection bias among schools uses the schoolwide distribution of MAP Performance Levels to calculate a school-level MAP Performance Index score. By the same token, the analysis of selection bias among teachers uses the percentage distribution of MAP Performance Levels for specific classrooms (and by extension, teachers) to calculate the MAP Performance Index. As the MAP Performance Index is an aggregate measure, it is not appropriate to discuss these differences in terms of lower levels of measurement. For example, any discussion of classroom-level differences based on the school-level MAP Performance Index would be inappropriate. Likewise, any discussion of student differences based on the classroom-level MAP Performance Index would be inappropriate.

The following analysis addresses the question of school selection: Were the schools participating in the FY02 eMINTS cohort statistically different from the non-eMINTS schools in the state? To answer this question, the eMINTS evaluation team has drawn together a sample of 39 non-eMINTS schools, based on available DESE core data for the 2002-2003 school year, to use as a comparison group for the FY02 eMINTS schools. This selection is a random, unstratified sample of non-eMINTS elementary schools. Schools are compared according to common administrative data available from the state school core data archive and the MAP test archive. The comparison has two parts: a comparison of aggregate building-level demographic characteristics and a comparison of performance on the MAP tests. Both of these comparisons suggest that the FY02 eMINTS schools did not differ from the matched sample of non-eMINTS schools.

Table 1
Percentage Distribution of School Location: Sample and eMINTS Schools

	Sample Schools	eMINTS FY02 Schools	All Schools
Central City	23.1	13.2	18.2
Suburban	35.9	15.8	26.0
Town	7.7	13.2	10.4
Rural	33.3	57.9	45.5
All Schools	100.0	100.0	100.0
Number of Schools	39	38	77
P-Value	0.0674		

Differences in Building Characteristics

Tables 1 and 2 present results for the comparison of sample schools and FY02 eMINTS schools on school location, student enrollment, percentage of minority enrollment and percentage of students eligible for the Free and Reduced Lunch program. Table 1 presents results by school location. Since school location does not change from year to year, only the results for the 2002-2003 school year are presented. Student populations can change substantially from year to year, so Table 2 reports statistics from two school years, 2000-2001 (the year the FY02 schools applied to participate in the eMINTS program) and 2002-2003.

Table 1 presents schools categorized by the urban and rural location codes used by the National Center of Education Statistics (NCES). No statistically significant differences between the sample schools and the FY02 eMINTS schools are apparent.

Table 2 presents statistics for student enrollment, the percentage of minority enrollment and the percentage of students eligible for the Free and Reduced Lunch program. Again, no statistically significant differences on these variables are apparent.

Differences in Student Performance at the School Level

Figures 1 and 2 and Tables 3 and 4 present MAP-score differences for the FY02 eMINTS schools and the sample schools. These results are expressed in terms of the MAP Performance Index and control for two important student-population variables: the percentage of low-income students enrolled in the school and the percentage of minority students in the school. These two variables are widely held to be the most important non-educational factors influencing school-level performance (see for example, Bernstein, J., & Rothstein, R., 1998).

The following results are studentized residuals taken from a school-level regression model controlling for the percentage of low-income students enrolled in the school and the percentage of minority students in the school. The results show where the set of FY02 eMINTS schools and sample schools scored relative to where one would expect them to

score given their school populations. As studentized residuals, these results express their predicted values in standard error units.

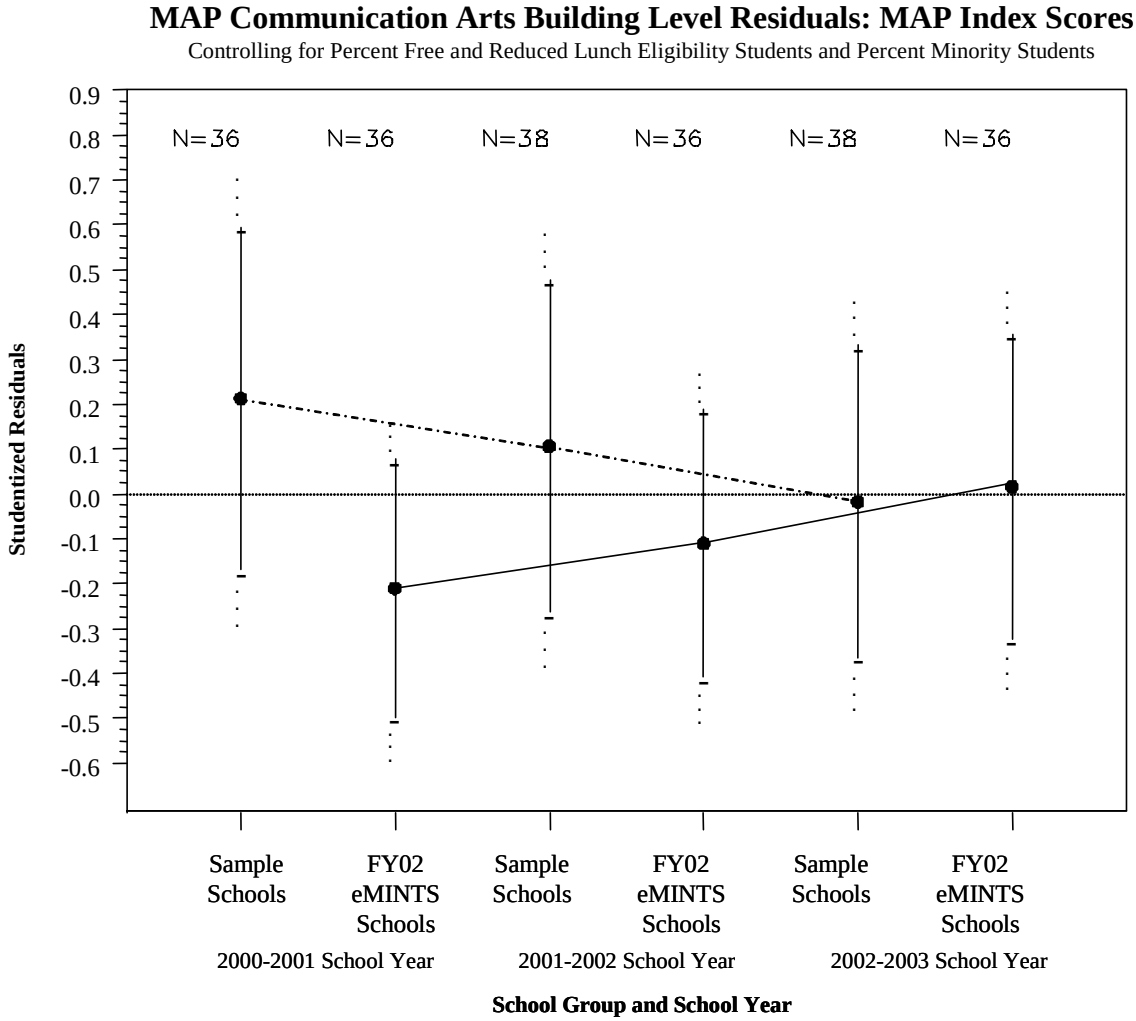
The retrospective nature of this analysis now has relevance. The figures and tables that follow present school-level residuals for the sample groups over three school years, from 2000-2001 through 2002-2003. These three years cover the FY02 eMINTS schools' participation in the program: 2000-2001 was the year the FY02 eMINTS schools applied to participate in the program; 2001-2002 was their first year of eMINTS professional development; and 2002-2003 was their final year of eMINTS professional development. During this period, several schools represented in the figures and tables that follow opened, others changed their grade configurations and still others decided to forgo the optional MAP tests (science in third grade and social studies in fourth grade). Consequently, the figures and tables that follow do not represent the full component of 78 (39 eMINTS and 39 sample) schools.[2]

[2] A full accounting of the variation among the set of 78 schools is available from the eMINTS evaluation project.

Table 2
Distribution of Student-Population Characteristics: Sample and eMINTS Schools
(Summary Statistics)

	Number of Schools	Mean	Standard Deviation	Lower 95% Confidence Limit for Mean	Upper 95% Confidence Limit for Mean
Building Size					
2000-2001 School Year					
Sample Schools	39	345.00	173.27	288.83	401.17
eMINTS FY02 Schools	37	347.27	151.86	296.64	397.90
All Schools	76	346.11	162.12	309.06	383.15
P-Value	0.9518				
2002-2003 School Year					
Sample Schools	39	331.69	157.85	280.52	382.86
eMINTS FY02 Schools	38	331.21	149.86	281.95	380.47
All Schools	77	331.45	152.95	296.74	366.17
P-Value	0.9891				
Percentage of Minority Students					
2000-2001 School Year					
Sample Schools	39	20.74	29.25	11.26	30.23
eMINTS FY02 Schools	37	12.99	27.42	3.85	22.14
All Schools	76	16.97	28.45	10.47	23.47
P-Value	0.2379				
2002-2003 School Year					
Sample Schools	39	21.88	29.15	12.43	31.34
eMINTS FY02 Schools	38	13.49	27.26	4.53	22.45
All Schools	77	17.74	28.36	11.30	24.18
P-Value	0.1959				
Percentage of Free and Reduced Lunch Students					
2000-2001 School Year					
Sample Schools	39	44.98	26.72	36.32	53.64
eMINTS FY02 Schools	37	50.43	19.09	44.06	56.79
All Schools	76	47.63	23.33	42.30	52.96
P-Value	0.3081				
2002-2003 School Year					
Sample Schools	39	48.30	26.72	39.63	56.96
eMINTS FY02 Schools	38	53.31	17.94	47.42	59.21
All Schools	77	50.77	22.81	45.60	55.95
P-Value	0.3357				

Figure 1
MAP Communication Arts Results by Sample Group (2000-2001 to 2002-2003)
Studentized Residuals

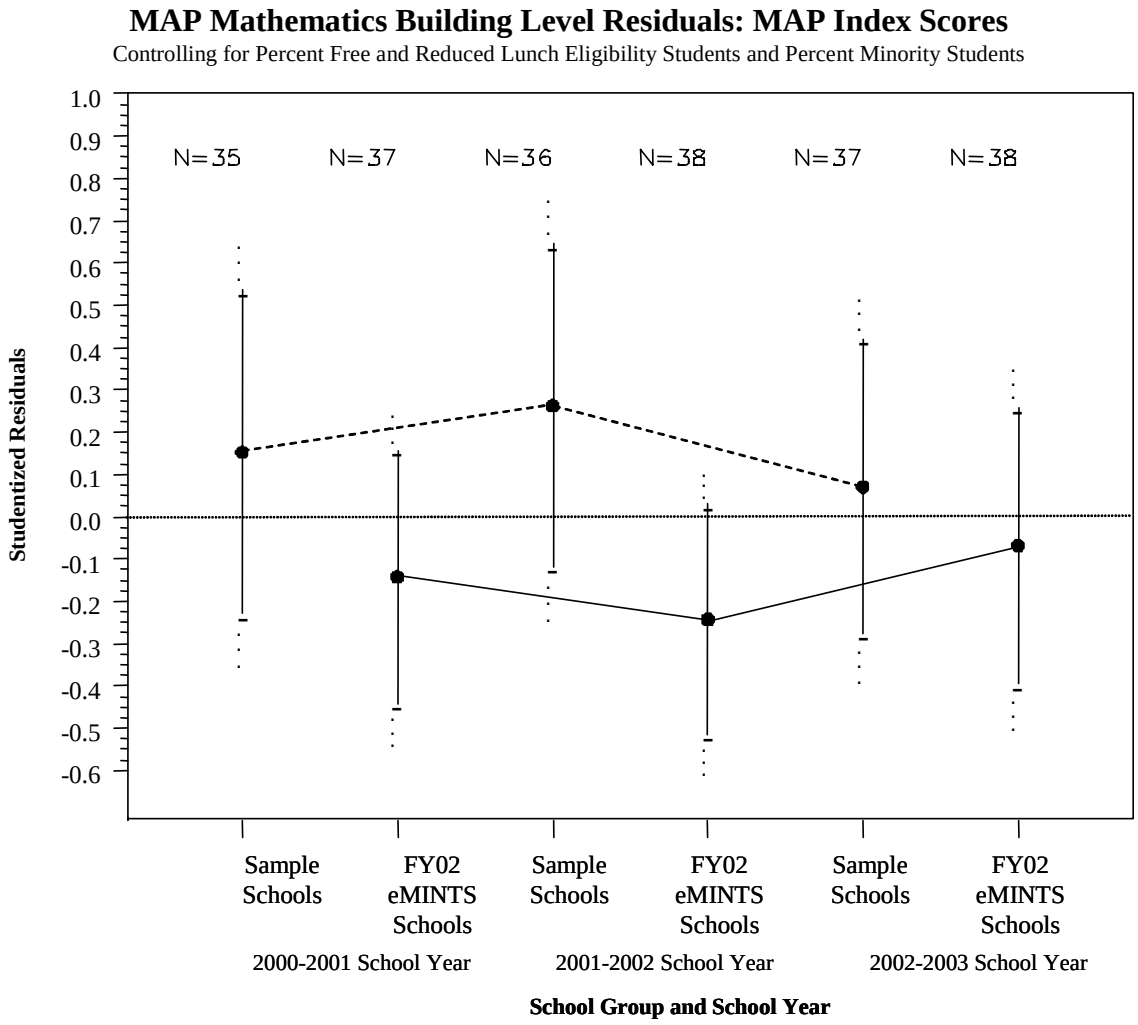


Despite the slight variation in the number of schools across the period, Figures 1 and 2 and Tables 3 and 4 show few statistically significant differences between the FY02 eMINTS cohort and the sample schools. No statistically significant differences exist on the communication arts test, and only one statistically significant difference appears for the 2001-2002 mathematics test, with the FY02 eMINTS schools scoring significantly lower than the sample schools.

Table 3
MAP Communication Arts Results by Sample Group (2000-2001 to 2002-2003)
Studentized Residuals, Summary Statistics

Student Enrollment	Number of Schools	Mean	Standard Deviation	Lower 95% Confidence Limit for Mean	Upper 95% Confidence Limit for Mean
2000-2001 School Year					
Sample Schools	36	0.21	1.13	-0.17	0.60
eMINTS Schools	36	-0.21	0.85	-0.50	0.08
All Schools, 2000-2001 School Year	72	0.00	1.02	-0.24	0.24
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		-0.42	0.0771		
2001-2002 School Year					
Sample Schools	38	0.11	1.13	-0.26	0.48
eMINTS Schools	36	-0.11	0.88	-0.41	0.19
All Schools, 2001-2002 School Year	74	0.00	1.02	-0.23	0.24
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		-0.22	0.3611		
2002-2003 School Year					
Sample Schools	38	-0.02	1.06	-0.37	0.33
eMINTS Schools	36	0.02	1.00	-0.32	0.36
All Schools, 2002-2003 School Year	74	0.00	1.03	-0.24	0.24
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		0.03	0.8904		

Figure 2
MAP Mathematics Results by Sample Group (2000-2001 to 2002-2003)
Studentized Residuals



These results, coupled with the analysis of school-background characteristics, suggest that no selection bias exists at the school level. According to the available data, the FY02 eMINTS schools do not differ statistically from a matched set of non-eMINTS elementary schools.

Table 4
MAP Mathematics Results by Sample Group (2000-2001 to 2002-2003)
Studentized Residuals, Summary Statistics

Student Enrollment	Number of Schools	Mean	Standard Deviation	Lower 95% Confidence Limit for Mean	Upper 95% Confidence Limit for Mean
2000-2001 School Year					
Sample Schools	35	0.15	1.11	-0.23	0.53
eMINTS Schools	37	-0.14	0.90	-0.44	0.16
All Schools, 2000-2001 School Year	72	0.00	1.01	-0.24	0.24
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		-0.29	0.2184		
2001-2002 School Year					
Sample Schools	36	0.26	1.13	-0.12	0.64
eMINTS Schools	38	-0.24	0.83	-0.52	0.03
All Schools, 2001-2002 School Year	74	0.00	1.01	-0.23	0.24
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		-0.51	0.0306		
2002-2003 School Year					
Sample Schools	37	0.07	1.04	-0.28	0.42
eMINTS Schools	38	-0.07	0.99	-0.39	0.26
All Schools, 2002-2003 School Year	75	0.00	1.01	-0.23	0.23
Differences in Means			P-Value		
eMINTS Schools vs. Sample Schools		-0.14	0.5551		

These results suggest that, among the schools in the FY02 cohort, the eMINTS schools did not perform any better than a random sample of schools statewide. In the one case where a statistically significant difference does exist, on the MAP Mathematics test for the 2001-2002 school year, the eMINTS schools scored lower on the MAP Performance Index than the non-eMINTS schools,

Assessing Selection Bias Among eMINTS Teachers and Students

The analytical tools available for assessing selection biases among schools are limited by a lack of complete data about all the individual schools that apply to be in an eMINTS program cohort in a given year. For example, complete information about the applicants for a cohort is not part of the data available to the eMINTS evaluation team. However, more data is available regarding classrooms, teachers and students. Administrative data about all the teachers in an eMINTS school is available through the state core data archive and equivalent data about all the students in an "eMINTS grade" is available from the MAP data archive. While neither data source contains all of the fields one might like to have in assessing selection (for example, there are no direct measure of teacher quality or of a student's prior-year test performance), the eMINTS evaluation team can use these data sources to estimate the effect of the nonrandom selection of teachers into the eMINTS program and of students into eMINTS classrooms. This section describes a methodology for this estimation and presents results for the FY02 cohort of eMINTS schools.

Estimating the Effect of Multilevel Selection

Statistical models for estimating sample selection bias have existed since the 1980s (for a review, see Fu, Winship and Mare, 2004). These models have proposed accounting for nonrandom selection by using a two-step method: first estimating the probability of selection into the treatment group and then including this estimator as an independent variable in a second model on the outcome measure. Through this approach, known as the "Heckman Method," after its developer, James Heckman, a researcher can determine both whether or not a selection bias exists—by assessing the statistical significance of the selection estimator—and the effect the nonrandom selection has on the outcome. Such models and estimators are common in a variety of econometrics contexts: estimating the effect of job-training programs on respondent income (Heckman and Hotz, 1989), the effects of health-reform policies on physician workload and prescription costs (Moatti, Paraponaris, Protopopescu and Verger, 2004), the effects of program participation on racial inequality (Mare and Winship, 1984). However, these methods have not been used extensively in educational evaluation[3].

The relative lack of work on evaluating selection bias in educational evaluation is due in part to two common characteristics of educational evaluations: the hierarchal nature of classrooms and a lack of sufficient data to estimate these types of models. The first is an obvious situation. When the focus is on differences in student performance, one must account for the clustering of students in classrooms and schools. Accounting for this clustering is a common motivation behind the development of hierarchal linear models

[3] One exception to this is the work of Robert Bifulco (2002).

(HLM), which explicitly account for group and classroom-level differences on individual achievement scores (Raudenbush and Bryk, 2003). The current analysis of teacher and student selection bias uses the Heckman Method in an HLM context.

The second issue, finding the data required to estimate the effects of selection bias, is more serious. In his analysis of whole-school reform efforts among New York City schools, Bifulco (2002) argues that multiple years of student-performance data are needed to understand the effects of a school's participation in a whole-school reform effort. He also recognizes that this data is not always available and provides a framework for using instrumental variables to help account for self-selection.

While the current analysis focuses on classrooms rather than schools, the data limitations that Bifulco describes are certainly relevant. All the available data is taken from administrative records (demographic data about students, basic certification and educational-attainment data about teachers and so forth). The most important variables, indicators for the actual selection process, are not available: measures of prior student performance, measures of prior teacher performance and so on. In addition, the student-performance data is available for a single year. For example, the third-grade students who took the MAP test in communication arts in the spring of 2003 will not test again in communication arts until the implementation of statewide annual testing in spring of 2006[4]. Under the current testing structure, the multiple years of student data that Bifulco argues are necessary to assess nonrandom selection bias do not exist. Consequently, one is left with an instrumental variable approach to understanding selection.

An Approach to Modeling Selection

As previously discussed, two selection processes influence the outcomes associated with eMINTS enrollment: the selection of teachers to participate in the eMINTS professional-development program and the selection of students into eMINTS classrooms. The process of modeling the effects of this selection process requires the estimation of three equations. The first equation predicts the selection of teachers and produces an instrumental variable representing the probability that any given teacher was selected to become an eMINTS teacher. The second equation predicts the selection of students and produces an instrumental variable representing the probability that any given student was enrolled in an eMINTS classroom. The third model combines the two instrumental variables in a HLM model to assess whether or not either selection process influences student MAP scores.

[4] Under the previous MAP testing schedule, the one that existed before the passage of the No Child Left Behind act, these students would not test again in communication arts until the spring of 2007, when they were in seventh grade.

These models have the following forms:

1. The Teacher-Selection Model:

$$\begin{aligned} pr(eMINTS\ Teacher) = \beta_0 + \\ \beta_1(Teacher\ has\ Masters\ Degree, 2000-2001) \\ \beta_2(Number\ of\ Subject\ Area\ Certificates, 2000-2001) + \\ \beta_3(Years\ Teaching\ Experience, 2000-2001) + u_j \end{aligned} \quad (1)$$

Following the conventions of the Heckman model, this model is a probit model predicting selection into the eMINTS program. This model produces a variable named the "Teacher-Selection Factor," or TS_j, which is a transformation of the predicted value of the probit model using the formula of the first derivative of the Inverse Mills Ratio (Stolzenberg and Relles, 1997).

All these independent variables are taken from the core data archive maintained by DESE. The values of each variable are as of the 2000-2001 school year, the year that the FY02 schools applied to the eMINTS program. These variables measure teacher educational attainment, general teacher expertise and teacher experience.

These measures are not direct measures of teacher quality, however. Given the limitations of the available teacher data, which favors administrative credentials over more impressionistic assessments of teacher quality or ability, it is not possible to determine the existing level of ability or performance for any teacher in the FY02 schools. Nevertheless, the existing model does provide some understanding of the differences between eMINTS and non-MINTS teachers.

2. The Student-Selection Model:

$$\begin{aligned} pr(eMINTS\ Student = 1) = \beta_0 + \\ \beta_1(Black\ Student) + \\ \beta_2(Other\ Race\ Student) + \\ \beta_3(Low-Income\ Student) + \\ \beta_4(Special\ Education\ Student) + r_i \end{aligned} \quad (2)$$

This model is also a probit model. Here, the data comes from the 2003 MAP tests. Each independent variable is taken from the student information accompanying the MAP-test record. This model produces a "Student-Selection Factor," or SS_{ij}. A separate factor is estimated for each MAP test. This factor is derived in the same manner as the Teacher-Selection Factor.

All these independent variables are dummy variables; the value 1 indicates that a student falls into a category and the value 0 indicates that the student does not fall into that

category. The "Low-Income Student" variable is derived by a student's eligibility for the Free and Reduced Lunch program. The "Special Education Student" variable indicates whether or not a student has an active Individual Education Plan (IEP). A student's diagnosis or the severity of a special-education student's disability cannot be determined from MAP-test records.

This model does not fully capture all the effects one would like. For example, no variable accounts for prior student ability. The 2003 MAP score is the first standardized test score for elementary students in each subject. Consequently, it is not possible to determine whether or not students selected for eMINTS classrooms performed any differently than other students in the years before the 2003 MAP test. Likewise, this model cannot account for many other unknown factors that may have influenced selection into an eMINTS classroom.

3. The Combined-Selection Model

The final selection model combines the Teacher- and Student-Selection Models into a multilevel model to determine the effect of nonrandom selection into an eMINTS classroom on student MAP scores.

The Student-Selection Model:

$$y_{ij} = \beta_{0j} + \beta_{1j} SS_{ij} + r_{ij} \quad (3.1)$$

$$\beta_{1j} = \gamma_{1j}$$

The Teacher-Selection Model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} TS_j + u_{0j} \quad (3.2)$$

The Combined-Selection Model:

$$y_{ij} = \gamma_{00} + \gamma_{01} TS_j + \gamma_{1j} SS_{ij} + u_{0j} + r_{ij} \quad (3.3)$$

The Combined-Selection Model, numbered 3.3, is a random-intercept model. This model assumes that the effect of teacher selection impacts all students in a classroom in the same manner, for example, by changing the intercept in the student-selection equation. This model also assumes that the impact of the Student-Selection Factor is the same across schools. If more than 39 schools participated in the FY02 eMINTS cohort, there would be enough school-level variation to estimate a three-level model to account for these differences.[5]

[5] This model also assumes that teacher and student selection processes are independent of each other. If there were evidence to suggest that this is not the case, it would be possible to add a cross-level interaction term to account for the effect of teacher selection on student selection.

Table 5
Descriptive Statistics for Teachers in FY02 eMINTS Schools
2000-2001 School Year: Grade Level Taught and Educational Attainment

	eMINTS Teacher		
	No	Yes	All Teachers
Grade Level Taught in 2000-2001 School Year			
2	26.5	2.4	21.6
3	25.6	36.5	27.8
4	24.4	56.5	30.8
5	23.5	4.7	19.8
Total	100.0	100.0	100.0
Masters Degree in 2000-2001 School Year			
No	68.2	77.6	70.1
Yes	31.8	22.4	29.9
Total	100.0	100.0	100.0
P-Value	0.0900		
Number of Teachers	340	85	425

Estimates of Selection

The next section presents the results from estimating these models. The Teacher-Selection Model is estimated once, for all eMINTS teachers, while the Student-Selection Model is estimated for the communication arts test and the mathematics test. The process of estimating the Teacher- and Student-Selection Models suggests some surprising differences among teachers and students, but the estimation of the combined models suggests that the selection biases do not impact student performance. This finding leads to a cautious conclusion that the MAP-score differences observed between students in eMINTS and non-eMINTS classrooms are not due to unequal selection, but *are* due to the impacts of the eMINTS program.

Results from the Teacher-Selection Model

Tables 5 and 6 present descriptive statistics for the variables in the Teacher-Selection Model. The data is taken from the 2000-2001 school year, the year the FY02 schools applied to participate in the eMINTS program. The 85 teachers selected to participate in the eMINTS professional-development program were chosen from 425 teachers in grades 2 through 5. Table 5 shows that teachers' overall educational attainment did not influence whether or not they were selected to participate in the eMINTS program.

Table 6
Descriptive Statistics for Teachers in FY02 eMINTS Schools, 2000-2001 School Year: Number of Subject-Area Certificates and Years of Teaching Experience

Number of Subject-Area Certificates
Summary Statistics

Level	Number of Observations	Mean	Standard Deviation	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Minimum	Maximum
eMINTS Teacher	85	1.64	0.80	1.46	1.81	1	4
Non-eMINTS Teacher	340	1.75	0.96	1.65	1.85	1	7
Total	425	1.73	0.93	1.64	1.82	1	7
			P-value				
Difference in Means		-0.11	0.2580				

Years of Teaching Experience
Summary Statistics

Level	Observations	Mean	Standard Deviation	Lower 95% Confidence Interval	Upper 95% Confidence Interval	Minimum	Maximum
eMINTS Teacher	85	9.60	7.64	7.95	11.25	0	27
Non-eMINTS Teacher	340	12.92	10.02	11.85	13.99	0	39
Total	425	12.26	9.67	11.33	13.18	0	39
			P-Value				
Difference in Means		-3.32	0.0010				

Table 6 shows the summary statistics and t-tests for the two continuous teacher variables: the number of subject-area certificates teachers earned by the 2000-2001 school year and the number of years of teaching experience teachers had in 2000-2001. No significant differences exist between eMINTS and non-eMINTS teachers in the number of certificates earned at the time of school application to the eMINTS program. However, eMINTS teachers have significantly fewer years of teaching experience. On average, eMINTS teachers have 3.32 fewer years of experience compared to their non-eMINTS peers. This suggests, contrary to the assumption that eMINTS teachers are "better" teachers than their non-eMINTS peers, that eMINTS teachers are less experienced than other teachers.

Table 7 presents the results for the Teacher-Selection Model itself. As expected, only the variable regarding the years of teaching experience is significant in predicting the selection of eMINTS teachers. The sign of this coefficient is negative, supporting the finding that eMINTS teachers have fewer years of experience than non-eMINTS teachers. However, this model is not a very strong predictor of selection. The final model has a pseudo-R^2 of 0.038[6]. The low value of this statistic suggests that none of the available variables are good predictors of selection.

[6] This measure is the "Adjusted Pseudo-R^2" proposed by Ben-Akiva and Lerman (1985:167). It accounts for the number of independent variables entered into the model. See Long (1997:104).

Table 7
Teacher-Selection Model for FY02 eMINTS Teachers

	Initial Model			Final Model		
	Coefficient	Standard Error	P-Value	Coefficient	Standard Error	P-Value
Intercept	-0.4533	0.1744	0.0094	-0.5863	0.1096	<0.0001
Masters Degree	-0.1049	0.1722	0.5422			
Number of Subject-Area Certificates	-0.0739	0.0818	0.3663			
Years of Teaching Experience	-0.0207	0.0083	0.0126	-0.0225	0.0078	0.0039
Log Likelihood	-207.606			-208.311		
Pseudo R-Square	0.032			0.038		

Results from the Student-Selection Models

Assessing patterns of student selection requires the estimation of one model for each test: communication arts and mathematics. Both of these models begin with the same set of student predictors: variables indicating student race, student poverty (as indicated by eligibility for the Free and Reduced Lunch program), and whether or not the student was receiving special-education services (as indicated by an active IEP at the time of the test administration). As seen in the analysis of FY02 MAP scores (Bickford, 2004) students identified as eligible for the Free and Reduced lunch program and students with active IEPs generally scored lower than other students.

As one of the predictors in these models is student race, it is necessary to remove four eMINTS schools from the analysis. These schools have relatively large populations of Black students. In two schools, all the students in grades 3 and 4 are Black and, in the other schools removed, more than 46% of the students in grades 3 and 4 are Black. Including these schools in the selection model makes the Black student variable statistically significant, suggesting that Black students are more likely to be enrolled in eMINTS classrooms than other students. However, this conclusion is incorrect for schools where more than 46% of the students in the eMINTS grades are Black. This selection has the effect of reducing the number of schools in the analysis to 35 and reducing the number of students in the analysis to 937 in grade 3 and 1634 in grade 4.

Table 8
Descriptive Statistics for Students in FY02 Communication Arts Test:
Student Race, Poverty Status and IEP Status

	eMINTS Classroom		
	No	Yes	All Students
Student Race			
White	95.1	93.3	94.1
Black	2.8	1.7	2.2
Other	2.1	5.0	3.7
All Students	100.0	100.0	100.0
P-Value	0.0224		
Student Eligible for Free and Reduced Lunch Program			
No	49.1	50.0	49.6
Yes	50.9	50.0	50.4
All Students	100.0	100.0	100.0
P-Value	0.7915		
Special-Education Student			
No	86.0	85.4	85.7
Yes	14.0	14.6	14.3
All Students	100.0	100.0	100.0
P-Value	0.7995		
Number of Students	401	536	937

Table 8 shows the percentage distribution by eMINTS status of the third-grade students in the communication arts data set. One significant difference exists between eMINTS and non-eMINTS students. Students classified in the Other race category, for example, students of American Indian, Asian and Hispanic origin, were more likely to be enrolled in eMINTS classrooms than either White or Black students. No differences exist for low-income students or special-education students. This suggests, given the available data, that eMINTS and non-eMINTS students did not differ from one another.

Table 9 shows the Student-Selection Model for the communication arts test. The results confirm the findings in Table 8. The only significant variable is whether or not a student is categorized in the Other race category. The probability that a student in this category is enrolled in an eMINTS classroom is 55% higher than the probability that a White student is enrolled in an eMINTS classroom. It must be noted that there are few third-graders in the Other race category. Overall, this group of students is less than 4% of all third-graders in the FY02 eMINTS schools.

The results from Tables 7 and 9 were used to calculate the Teacher- and Student-Selection Factors used to estimate the overall impact of nonrandom selection on the communication arts test scores. Table 10 presents the descriptive statistics for these two factors, and Table 11 presents the HLM model estimating the effects of these factors.

Table 11 shows that neither of the selection factors have a significant impact on the communication arts test score. This finding suggests that the indefinable grounds for teacher and student selection did not influence the test scores of students enrolled in eMINTS classrooms. Rather, it supports the claim that the six-point difference on the communication arts test between students enrolled in eMINTS classrooms and students not enrolled in eMINTS classrooms is due to the eMINTS teachers' application of the instructional methods presented in the eMINTS professional-development program.

Table 9
Student-Selection Model for FY02 MAP Communication Arts Test

	Initial Model			Final Model		
	Coefficient	Standard Error	P-Value	Coefficient	Standard Error	P-Value
Intercept	0.178	0.0603	0.0031	0.1696	0.0426	<0.0001
Black Student	-0.2978	0.2848	0.2958			
Other Race Student	0.5613	0.2425	0.0207	0.5519	0.2404	0.0217
Student Eligible for Free and Reduced Lunch Program	-0.0279	0.0851	0.7431			
Special-Education Student	0.0811	0.1208	0.5021			
Log Likelihood	-616.952			-617.775		
Pseudo R-Square	0.029			0.033		

Table 10
Descriptive Statistics for Teacher- and Student-Selection Factors
FY02 MAP Communication Arts Test

	Number of Classes	Number of Students	Mean	Standard Deviation	Minimum	Maximum
Teacher-Selection Factor	29	536	1.39	0.15	1.22	1.68
Student-Selection Factor	29	523	0.68	0.06	0.40	0.69

Table 11
HLM Selection Model, FY02 MAP Communication Arts Test

	Initial Model				Final Model			
	Coefficient	Standard Error	Df	P-Value	Coefficient	Standard Error	Df	P-Value
Intercept	643.14	1.79	28	<0.0001	637.42	21.7657	27	<0.0001
Teacher-Selection Factor					9.6858	12.1029	492	0.4239
Student-Selection Factor					-11.4212	20.956	492	0.5860
Model P-Value	0.0036				0.0042			
Residual Variance	801.68				809.73			
% Improvement					1.00			
Number of Students	535				522			
Number of Classrooms	29				29			

Tables 12 through 15 repeat the selection analysis for the FY02 MAP Mathematics test. This test was administered to students in the fourth grade. The major difference between these results and the results for the communication arts test is the lack of significant Student-Selection Factors. In the third grade, students in the Other racial category were more likely to be enrolled in an eMINTS classroom than other students. In the fourth grade, no student factors are statistically significant. The probit model in Table 13 confirms this finding. Consequently, it is not possible to estimate a Student-Selection Factor.

Table 14 presents the descriptive statistics for the Teacher-Selection Factor, and Table 15 presents the HLM Selection Model. As it was with the communication arts model, the Teacher-Selection Factor does not have a significant impact on the MAP Mathematics test score. This finding, again, suggests that the overall differences between students enrolled eMINTS classrooms and students not enrolled in eMINTS classrooms is due more to the effect of the eMINTS program than to nonrandom selection.

Summary and Conclusions

This report has been motivated by the development of a body of literature and policy that seeks to make educational evaluation studies and analysis more experimental. Advocates for what has become known as "scientifically based research" have embraced a simplistic model for study design and participant selection. This model forms the basis of a spirited debate over the quality and validity of educational research findings.

The eMINTS evaluation has been caught up in this debate since the publication of its first policy briefs in 2000. The current report seeks to describe the programmatic constraints that complicate the simplistic model of randomized selection. As previously discussed, the eMINTS selection process has at least three levels, none of which can be easily randomized. Using the experimental model favored by advocates of "scientifically based research," the lack of random selection would render the general results of the eMINTS evaluation invalid. However, as evaluation practitioners and educators, we would do well to be more cautious.

The current report has taken the "scientifically based research" critique of eMINTS seriously. The report has outlined some of the context behind the design decisions the eMINTS evaluation project has made. It has also proposed a methodology for testing whether or not the sample-selection biases introduced at each level influence the overall MAP-score difference between students enrolled in eMINTS classrooms and students not enrolled in eMINTS classrooms.

Table 12
Descriptive Statistics for Students in FY02 MAP Mathematics Test:
Student Race, Poverty Status and IEP Status

	eMINTS Classroom		
	No	Yes	All Students
Student Race			
White	94.7	94.7	94.7
Black	2.4	2.6	2.5
Other	3.0	2.7	2.8
All Students	100.0	100.0	100.0
P-Value	0.7652		
Student Eligible for Free and Reduced Lunch Program			
No	55.1	55.2	55.1
Yes	44.9	44.8	44.9
All Students	100.0	100.0	100.0
P-Value	0.9899		
Special-Education Student			
No	86.3	85.7	85.9
Yes	13.7	14.3	14.1
All Students	100.0	100.0	100.0
P-Value	0.7140		
Number of Students	693	941	1634

Table 13
Student-Selection Model for FY02 MAP Mathematics Test

	Initial Model		
	Coefficient	Standard Error	P-Value
Intercept	0.1945	0.0433	<0.0001
Black Student	0.0614	0.2042	0.7636
Other Race Student	-0.0498	0.1908	0.7941
Student Eligible for Free and Reduced Lunch Program	-0.0132	0.0647	0.8387
Special-Education Student	0.0428	0.092	0.6416
Log Likelihood	-1088.985		
Pseudo R-Square	0.019		

Table 14
Descriptive Statistics for Teacher- and Student-Selection Factors, FY02 MAP Mathematics Test

	Number of Classes	Number of Students	Mean	Standard Deviation	Minimum	Maximum
Teacher-Selection Factor	48	941	1.36	0.13	1.20	1.68

Table 15
HLM Selection Model, FY02 MAP Mathematics Test

	Initial Model				Final Model			
	Coefficient	Standard Error	Df	P-Value	Coefficient	Standard Error	Df	P-Value
Intercept	645.29	2.07	47	<0.0001	639.80	22.93	46	<0.0001
Teacher-Selection Factor					4.02	16.73	884	0.8100
Model P-Value	<0.0001				<0.0001			
Residual Variance	1163.29				1163.15			
% Improvement					-0.01			
Number of Students	932				932			
Number of Classrooms	48				48			

The test score differences between students enrolled in eMINTS classrooms and their peers could be due to any number of factors including the differential selection of teachers and the differential enrollment of students. However, the analysis of selection bias in the FY02 cohort suggests, given the limitations of the available data, that this scenario is not the case. None of the selection factors estimated were statistically significant, a finding that supports the hypothesis that the score differences seen in eMINTS schools are due the influence of the eMINTS program.

The passage of the No Child Left Behind legislation began a period of change in state-sponsored testing and in the practice of educational evaluation. The development of understanding regarding the complexity of the eMINTS program, as well as the establishment of a methodology to estimate selection bias, comes as part of those changes. In the future, annual testing will provide better data to estimate the impact of student selection. Understanding of the teacher-selection process will also improve. As its understanding of these processes improves, the eMINTS evaluation team will be able to better describe the impact of the eMINTS program on education in Missouri.

References

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Applications to Travel Demand*. Cambridge, MA: MIT Press.

Bernstein, J., & Rothstein, R. (1998, November/December). The black-white test score gap. *The American Prospect*,*9*.

Bickford, A. (2004). *Analysis of 2003 MAP Results for eMINTS Students.* Columbia, Missouri: Office of Social and Economic Data Analysis. http://www.emints.org/evaluation/reports/map2003.pdf

Bifulco, R. (2002). Addressing Self-Selection Bias in Quasi-Experimental Evaluations of Whole-School Reform: A Comparison of Models. *Evaluation Review*, 26:5, 545-572.

eMINTS Evaluation Project (2001). *A General Typology of eMINTS Lessons*. Columbia, Missouri: Office of Social and Economic Data Analysis. http://www.emints.org/evaluation/reports/lesson-typology.pdf

Fu, V. K., Winship, C. and Mare, R. (2004). Sample Selection Bias Models. In M. Hardy and A. Bryman (Eds.), *Handbook of Data Analysis* (409-430). Thousand Oaks: Sage.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47:1, 153-161

Heckman, J. J. and Hotz, V. J. (1989). Choosing Among Alternative Nonexperimental Models for Estimating the Impact of Social Programs: The Case of Manpower Training. *Journal of the American Statistical Association*, 84:408, 862-874.

Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

Mare, R. and Winship, C. (1984). The Paradox of Lessening Racial Inequality Among Black Youth: Enrollment, Enlistment and Employment, 1964-1981. American Sociological Review, 49, 39-55.

Moatti, J., Paraponaris, A., Protopopescu, C., Verger, P. (2004). Testing for selection bias in a simultaneous equations model of general practitioners' workload and prescription costs. *Applied Health Economics and Health Policy*, 3:1. http://www.openmindjournals.com/health.html

National Research Council. (2002). *Scientific Research in Education*. Washington, DC: National Academy Press.

Raudenbush, S. W. and Bryk, A. S. (2003). *Hierarchal Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks: Sage.

Reckase, M. D. (2004). The Real World is More Complicated that We Would Like. *Journal of Educational and Behavioral Statistics*, 29:1, 117-120.

Rubin, D. B., Stuart E. A., and Zanutto, E. L. (2004). A Potential Outcomes View of Value-Added Assessment in Education. *Journal of Educational and Behavioral Statistics*, 29:1, 103-116.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin.

Stolzenberg, R. M. and Relles, D. A. (1997). Tools for Intuition about Sample Selection Bias and its Correction. *American Sociological Review*, 62:3, 494-507.

Tharp, S. B. (2003). *Classroom Climate, Instructional Practices and Effective Behavior Management in eMINTS Expansion Classrooms*. Columbia, Missouri: Office of Social and Economic Data Analysis. http://www.emints.org/evaluation/reports/expansion2.pdf

Tharp, S. B. (2004). *Classroom Climate, Instructional Practice and Mentorship Experience in the eMINTS Expansion Classrooms: A Two-Year Study*. Columbia, Missouri: Office of Social and Economic Data Analysis. http://www.emints.org/evaluation/reports/expansionclassroom-climate.pdf

What Works Clearinghouse. (2002). *WWC Study Review Standards*. http://www.w-w-c.org/reports/study_standards_final.pdf

Winship, C. and Mare, R. (1992). Models for Sample Selection Bias. *Annual Review of Sociology*, 18, 327-350.