MEMORANDUM

MATHEMATICA
Policy Research, Inc.

TO: Advisory Panel Members

FROM: Mark Dynarski, Margaret Honey, and Doug Levin

DATE: 1/6/2003
Edtech-006

SUBJECT: Considerations for Designing a Study of the Effectiveness of Educational Technology

---

Since the first advisory panel meeting in November 2002, the design team has had regular discussions and has continued its efforts to identify approaches the study design could use to identify promising applications for the national study. This memorandum summarizes our thinking on design approaches and suggests questions that would be useful to discuss for discussion during the January 2003 meeting. The memo discusses:

- Research questions that provide a framework for the study;
- Statistical design considerations that relate to statistical power and estimated minimum sample sizes necessary to answer the research questions;
- Candidate technology applications to study, with information in the accompanying packet of materials that describes the applications in more detail and provides some of the research findings that are provided by publishers on web sites or are published in research journals.

1. Research Questions

The No Child Left Behind Act of 2001 mandates a rigorous evaluation of the effectiveness of educational technology. It also provides research questions that should be addressed by the study. These questions include:

1. Can student academic outcome measures be improved through the use of technology applications?
2. Which conditions and practices are necessary and sufficient to realizing improvements in student academic outcomes?
3. Which conditions and practices support effective applications?

4. Which conditions and practices best support schools and teachers in also improving student technology literacy?

Formulating research designs to answer these questions is the main purpose of this design effort and the focus of meetings with the advisory panel. It is clear from examining the questions that the study has a broad mandate, and attempting to answer each question at the highest levels of rigor could require more resources than have been set aside for the effort (an issue we return to below). An important result of the November panel meeting was to place priorities on how a study would answer the various questions. Important points raised in the discussion were:

- Technology applications could be chosen based on a number of criteria, including prior evidence of effects or expert judgment about promise of effects, prevalence of use, feasibility for going to scale, to fill the gap between prevalence and what is known about effectiveness from existing research, or to respond to teacher or school needs.). In addition, several panel members encouraged the team to consider web-based research applications that are used in many middle schools and high schools and whose use continues to grow.

- A range of academic outcome measures should be considered. Critical outcomes include reading comprehension and mathematical skills for young learners (grades K-2), algebra or geometry comprehension for older learners (these are considered gateway courses for college preparation.

- Conditions and practices under which technology applications are effective were recognized as important but no specific approaches to studying them were formulated.

In addition, discussion after the meeting led to clarifying that the technology applications to focus on would be those that were being supported or that could be supported with funding from the Enhancing Education Through Technology Program or from Title I, which primarily operate at the K-12 grade levels. Applications for special education students, postsecondary students, and adult education students would not be considered. Their primary funding is through other legislation and studies of technology effectiveness in these domains could be supported by that legislation. Still at issue is whether the study should consider technology approaches to help students learn English as a second language, which is now included under No Child Left Behind.

Some ideas were discussed at the first meeting but further consideration suggests that the design effort should focus less attention on them. These include studying technology applications whose focus is to assist teachers in diagnosing and assessing particular academic competencies so that they can better focus their teaching; distance learning applications (including those related to home schooling and virtual charter schools), and ubiquitous technology applications whose purposes are to promote a better management or information

services within school districts, such as networking, e-mail, and the Office© suite. While the team believes these applications play a role in teaching and learning, we elected to focus on technology applications that would be used in the classroom by a student (and probably with a teacher present) to support teaching and learning. However, we recognize that this issue may benefit from further consideration and reflection.

Although these decisions have helped to narrow the scope of the study, other issues about technology applications still imply a relatively large study. Within the more narrowly defined scope, technology applications can be categorized along five dimensions:

1. Instructional approach: drill-and-practice, meaning based, whole-school/district/state
2. Instructional level: primary, secondary
3. Subject area: math, reading, science, social studies
4. Prevalence: how widely in use, whether use is expected to grow
5. Setting: existing applications or new implementations

In addition, the legislation requires that the study examines the conditions and practices that support each technology application--a requirement that further expands the scope of the study.

2. Statistical Power and Sample Size Considerations

The idea emerging from the first meeting that the study should examine technology applications that operate across a range of subjects and grade levels raises an important issue about sample sizes that would be needed to measure effects with adequate precision. The set-aside for the study in the No Child Left Behind act, while substantial (about $14 million), nonetheless places limits on what the study can accomplish. Acknowledging these limits at the outset helps to guide discussions about how many applications can be included, and how many districts and schools may need to be recruited for the study.

Several assumptions were made to arrive at sample size estimates. First, we assumed that random assignment would need to be done at the classroom level (or, alternatively, assignment would be at the school level if a school had only one classroom for a grade level). Because technology applications often involve classroom-level implementation, teacher training, and peer activities, assigning students individually to receive the application would seem infeasible at the outset. Assigning whole classrooms to receive an application, however, introduces “intra-cluster correlation,” which arises whenever grouping creates a set of individuals who are more similar to each other than to individuals outside the group. For example, a school that assigned students to classess according to their reading levels would create a positive intra-cluster correlation on reading test scores (as would be the case if the school grouped students on any outcome). The larger the intra-cluster correlation, the larger the number of clusters (i.e, classrooms) that are

needed to achieve the same level of statistical precision, holding the overall number of students in the sample fixed. Experience suggests that intra-cluster correlations ranging from 5 percent to 20 percent of total variance probably bound the true value. As will be shown below, a 20 percent intra-cluster correlation implies that almost the entire study budget would be needed to study the effects of *one* technology application.[1]

Second, assumptions about the costs of collecting data for classrooms and students are needed so that the total costs under various sample sizes can be estimated. Using experience from recent large-scale random assignment studies in schools, we developed two cost factors: each cluster was assumed to cost an additional $30,000 (for identifying, recruiting, and studying each cluster), and each student was assumed to cost an additional $1,000 (for baseline and two follow-up data collection efforts, including administering tests and possibly a teacher or parent survey). In practice these cost factors are affected by a variety of considerations—the size of the district, the degree of cooperation with the study, the nature of districts approval processes to carry out research studies, and so on—but the calculations easily can be done to reflect other assumptions as needed.

Third, we assumed we want to detect an effect size of *20 percent or greater*. An effect size is the change in the outcome generated by the treatment as a proportion of the outcome's standard deviation. For example, for a test that had a mean of 100 and a standard deviation of 15 (characteristic of nationally-normed intelligence tests), a 20 percent effect size would be an increase in the mean score from 100 to 103 (the 3-point increase is 20 percent of the standard deviation of 15). The 20 percent target effect size is arbitrary to some degree, but a reasonable goal for policy. Several well-known experimental studies, such as the Tennessee STAR experiment and the Early Head Start experiment, have achieved effect sizes of about this magnitude, and evidence surveyed by Murphy et al. (2002) suggests that technology applications can achieve these effect sizes.

If evidence indicates that technology applications had larger effect sizes, the design would call for smaller sample sizes. However, as will be shown below, substantial cost implications arise if technology applications are believed to have smaller effect sizes and if detecting these effect sizes is important for policy. We return to the issue of effect sizes in the concluding remarks.

[1]Estimates of intra-cluster correlations are not commonly available in research literature and the design team will be looking further into the issue of estimating intra-cluster correlations using public data sources or evaluation databases.

Figure 1
Minimum Detectable Effect Sizes
Under a Cost Constraint With Optimal Clusters

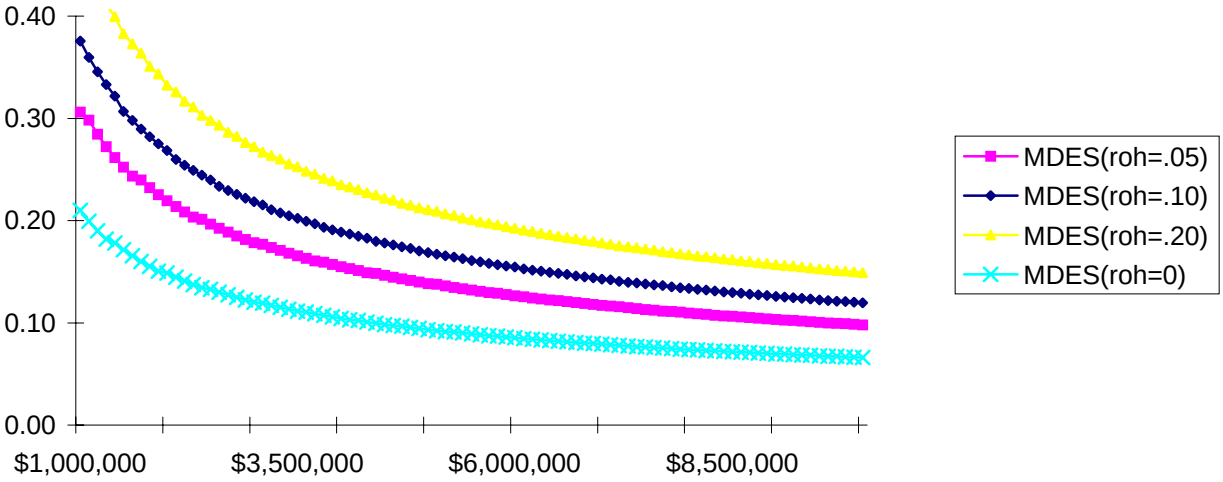

Figure 1 plots minimum detectable effect sizes when the number of students in a cluster is "optimized," meaning that the size of the clusters and the number of clusters is chosen to maximize the study's precision.[2] For fixed dollar expenditures (on the X axis), the figure shows minimum detectable effect sizes (the effect size at which a t-test of statistical significance has an 80 percent likelihood of indicating significance when the true effect size equals the minimum). Five observations based on the figure should be noted:

1. Minimum detectable effect sizes are lower when intra-cluster correlations are lower (intra-cluster correlations are termed "roh" factors—for "rates of homogeneity"—in the figure). For an expenditure of $3 million, for example, the minimum detectable effect size is 17.8 percent when the correlation is .05, 21.9 percent when it is .10, and

[2]Variance formulas for optimal designs under cost constraints are derived by Stephen Raudenbush, Statistical Analysis and Optimal Design for Cluster Randomized Trials, *Psychological Methods*, 3(2), 173-185, 1997. Raudenbush's formulas allow a covariate to reduce both within-variance and between-variance by different magnitudes, and in a related study, Bloom et al. (*Evaluation Review*, 1999) show that between-variance is reduced substantially when test scores are the outcome being studied and a baseline test score is available. The calculations here use a conservative approach and assume that a covariate reduces between and within variance by 20 percent. Larger reductions in between-variance would suggest smaller minimum detectable effects.

27.2 percent if it is .20. In other words, if we believe or estimate that the intra-cluster correlation is .10 or greater, we would need to set aside more than $3 million to have a reasonable chance of detecting a 20 percent effect size. Comparing the bottom-most curve in the figure where roh is assumed to be zero indicates how costly intra-cluster correlation is to study design. Only $1 million is needed to detect a 20 percent effect size when roh is zero.

2. Lower target effect sizes can require much greater resources. For example, assuming roh was .10 (the middle curve), detecting a 20 percent effect size requires $3 million and detecting a 15 percent effect size requires $6.3 million.

3. Optimal design calls for many classrooms. For example, assuming roh is .10 and a 20 percent effect size is desired, optimal design calls for the study to randomly assign and collect data in 78 classrooms, with a sample of 16 students in each classroom.[3] Including 78 classrooms suggests recruiting 20 to 40 schools (depending on school size), and even though one or two large urban districts possibly could encompass nearly the entire sample, the need for broader representativeness argues for including a range of districts in the effort, varying on whether they are urban or rural, high-poverty or low-poverty, and in different regions.

4. An important question that the study would likely face is about *differences* in effect sizes. For example, policymakers may want to know whether one application is more effective than a competing application. The results in Figure 1 are a caution about the study's ability to detect differences in effects. If two applications had effect sizes that differed by 10 percentage points (for example, one had an effect size of 25 percent and the other had an effect size of 15 percent), a huge and expensive sample would be required to have a reasonable chance of showing that the difference was statistically significant (if roh were .20, all $10 million would need to be spent on comparing the two applications). A design that is not based on randomly assigning large units is better suited for comparing applications. For example, designs that randomly assign students within a school to receive one application or the other may be more suitable for contrasting applications.

5. The same caution about the study's ability to detect differences applies to the study's ability to measure effect differences for the same application being implemented under differing conditions and practices. For example, an application could be implemented in classrooms whose teachers receive ample professional development and in classrooms whose teachers do not. The question of whether professional

[3] Even if classrooms contained more students (and most classrooms probably do), including additional students in a classroom in the study beyond the target of 16 reduces statistical precision, since the students would be included at the expense of having fewer classrooms in the study. However, from an operational perspective, collecting data for all students in a classroom may offer some economies.

development (or other classroom, school, or district factors) moderates effect sizes is statistically similar to the question of whether one application has larger effects than another. In both cases, the magnitudes and precision of the two measured effects are at issue. Large differences between precisely measured effects are more likely to be statistically significant. However, whether moderating factors will generate large differences cannot be presumed (and available literature does not offer much guidance about the size of the differences), and precision levels for effects that are based on subsamples will be less than what is shown in Figure 1.

The important conclusion from this analysis is that under the assumptions made here, *three substudies of technology applications would fit within available resources* (which at this point is assumed to be about $10 million for data collection). The scheme noted in the introduction to explore applications for early readers, later math learners, and high school students fits within the cost constraints. If other designs are identified that supported higher levels of precision, the range of the study could be expanded to include other possibilities such as distance learning. If later considerations suggest that the designs will be less powerful, fewer or smaller substudies would fit within the budget.

The above calculations do not consider the effects on statistical precision of district and school clustering factors. The calculations implicitly assume that classrooms are being randomly assigned rather than schools that contain the classrooms. If technology applications require a school-level implementation, the study would need to randomly assign schools rather than classrooms. The larger unit size of schools compared to classrooms, and the related possibility of larger intra-cluster correlations, may lead to studies that require more schools (and classrooms) to achieve precise measures of effects than what is shown in the figure. Likewise, considerations about district variability may lead to the study needing more classrooms. (If districts themselves add variance to measured effects, due to differences in district implementations and couterfactuals, for example, the design would need to include more districts to achieve the precision standard.) These considerations suggest that we view the estimates in Figure 1 is upper bounds for precision.[4]

The main source of the study design expense is in the clustering of students in classrooms, and approaches that reduce the effects of clustering would be useful to consider. For example, it may be possible depending on the application to assign

[4]Bloom *et al.* show using third-grade and sixth –grade test score data for reading and math from one district (Rochester, New York) that achieving a 20-percent minimum detectable effect size would require 40 schools, when a baseline test score is available as a covariate. These estimates suggest that a substudy target of 40 schools is reasonable.

students to classrooms, some that use the application and some that do not. Doing so would enhance statistical power. An issue would arise if the technology application affected teachers in treatment classrooms in ways that spilled over to teachers in control classrooms. Another approach would be to have the same teachers use two different approaches, one enhanced by technology, if they had different sections of a subject class such as algebra or geometry. Again, spillover of techniques needs to be considered, but the gain of precision makes such approaches well worth considering.

3. Candidate Technology Applications

An approach that fits within the statistical limits noted above would be to identify one or two representative or promising applications within the three categories of early readers (primary grades), later math learners (secondary grades), and high school learners. The accompanying packet abstracts information from developer web sites for emerging or popular applications, supplemented by research findings when they are available. The information is not exhaustive and does not list many applications. Information and reviews of many applications can be found on education web sites such as www.superkids.com, and Murphy *et al*. (2002) review the research findings in greater detail.[5]

- **Early reading**

Technology applications to support teaching and learning of reading are numerous and the literature studying the effects of the applications have suggested that their use leads to large reading gains.[6] A review of the research literature that accompanies many of the applications suggests the size of the claimed gains. The different designs used in the research and variations in how the results are reported suggest that some caution should be exercised in interpreting these findings. Also, some applications such as *Soliloquy Reading Assistant* and *Reading For Meaning* cite research to support their use but the research is published work by the National Reading Panel and others supporting the concepts underlying the application's design.

Accelerated Reader: A gain of 27 percentile points for low readers (small gains for other readers)

Waterford Early Reading Program: In Dallas, a gain of 13.4 normal-curve-equivalent points on the vocabulary test compared to a district gain of 0.7 points; in Norwalk, Connecticut, an effect

[5]See Murphy, R, *et al.*, E-Desk: A Review of Recent Evidence on the Effectiveness of Discrete Educational Software, April 2002, downloaded from www.sri.com/policy/ctl/pdfs/Task3DraftFinalReport3.pdf.

[6]Applications are numerous; for example, www.superkids.com alone provides reviews for forty reading applications and 10 math applications.

size of 0.61 for the full sample and 2.38 for lower-performing students; statistically significant improvements found in studies conducted in Newark, New Jersey, Hacienda-La Puente, California, and Prince George's County, Maryland.

Fast ForWord: "Students who train with Fast ForWord Language make language gains of 1-2 years in just 4-8 weeks." (Lessons are 100 minutes a day, five days a week during the 4-8 week span.) Other materials illustrate that Fast ForWord had effect sizes for Woodcock-Johnson test scores of about 15 to 20 percent for general students, and 50 percent for students at risk of reading failure.

WiggleWorks: Students using WiggleWorks increased a full grade level in reading compared to similar students.

- **Early Math**

  Building Blocks Effect sizes of 0.85 and 1.44 for number and geometry.

- **Secondary Math**

  Carnegie Tutor: On TIMSS test, a 20 percent effect size (Pittsburgh students), on problem-solving, a 44 percent effect size, similar results in other cities.

  I CAN LEARN: A 25 percent effect size on passing end-of-year algebra tests in Fort Worth, Texas, a 10 percentage-point higher score on the math portion of the Florida Comprehensive Assessment Test in Tampa, Florida.

  Geometer's Sketchpad: No evaluations were found.

  Geometric Supposer: No evaluations were found.

- **Secondary students, web-based instructional modules**

  Center for Improved Engineering and Science Education, Triarchic Enhancement: Small and insignificant differences in physics achievement scores

  Center for Children and Technology, JASON project "Most JASON (middle school) students acquired scientific inquiry and analytical skills, and outperformed non-JASON students based on the results of a pre- and post-inquiry test…"

Global Learning and Observation to Benefit the Environment (GLOBE) GLOBE students scored better on tests in earth science content.

Online communications, Center for Applied Special Technology: Middle school students using online communication scored statistically better on five of nine learning measures.

The review suggests that applications exist whose putative effect sizes are larger than the target effect size of 20 percent used here to estimate sample sizes. A recently completed and more thorough survey of research on the effectiveness of discrete technology applications (Murphy *et al.* 2002) also found effect sizes of 20 percent or greater. However, the same survey also cautions that the lack of rigorous studies using random-assignment or well-constructed comparison groups to measure effects leaves a wide range of uncertainty about effect sizes.

4. Issues for Discussion

Important issues need to be considered as the design team moves ahead toward concrete approaches for the study. The second advisory panel meeting is a useful juncture for further discussion of the issues.

a) Study applications that are or will be prevalent but about which research is limited or inconclusive

In the first advisory panel meeting, various panel members suggested that the study could identify applications that were anticipated to be prevalent during the next four to five years and for which research has not yet demonstrated compelling evidence of effectiveness. Directed web-based research was offered as an example. Other approaches discussed during the meeting included identify applications that served to promote equity and identifying applications that responded to expressed teacher and school needs. It would be useful to focus again on the issue of developing selection criteria. Combinations of the above criteria could be developed that looked at candidate applications from a variety of perspectives and arrayed their fit from the different perspectives. However, the importance placed on early reading in No Child Left Behind suggests that the study needs to select at least technology applications to support the teaching of reading.

b) Focusing on new or existing implementations

Also raised as an issue in the first meeting was whether the study should measure effects of technology applications as they are currently being used or measure effects in schools and districts that would be only beginning to implement them. The first approach puts the

application in the most favorable light by allowing implementation issues to have been worked out. The second approach strives for greater policy relevance because new expenditures would involve new implementations and measuring the effects in such cases provides a view of whether the return to investment is adequate. Identifying schools to implement a new technology application may pose difficulties for the recruiting effort, however. Also, considering the time frame of the study, the new applications that could be studied probably would need to be limited to those that can be implemented quickly.

c) Understanding the role of conditions and practices

A considerable portion of the first meeting was spent grappling with the relationship between conditions and practices and technology's effectiveness. The panel recognized the crucial role played by context factors such as teacher preparation and training in smoothly implementing technology and in promoting its effectiveness. However, the discussion also noted that some technology applications strive to have teachers not be part of the equation, which enables the applications to be implemented in a range of schools and settings with less concern about the training and experiences of the teachers. The sample size discussion plays a role in this issue, since it is clear that the ability to structure a design that rigorously measures the moderating effects of conditions and practices is limited. In particular, other approaches for understanding the moderating effects of conditions and practices would be useful to consider.

cc: Audrey Pendleton, Lara Hulsey, Roberto Agodini