The Mid-Atlantic Regional Educational Laboratory

The LSS Review

Laboratory for Student Success

a catalyst for merging research, policy, and practice

Volume 3 • Number 2 • December 2004

The Scientific Basis of Educational Productivity

Proceedings and Recommendations from a National Invitational Conference

Rena F. Subotnik, American Psychological Association, and Herbert J. Walberg, University of Illinois at Chicago

Beginners in many disciplines learn that correlation never proves causation, but sometimes, even in public health, correlation, mistaken for causation, becomes the basis for policy and great expenditures of public and private money. "True experiments" with random assignment to experimental and control groups hold a special place in the realm of scientific research. The results of such experiments, particularly when replicated under many, varied conditions, provide the most dependable basis for policy and practice, as clearly demonstrated and even required for definitive conclusions in agronomy and medicine.

The case for experiments is pressing in K–12 education, which lacks a strong foundation of causal research, particularly disciplined-based control group experiments and large-scale, well-controlled statistical studies. Given the strong consensus among policymakers about the need for improved academic performance on the part of our nation's students—as evidenced by the federal No Child Left Behind Act and more stringent state testing and accountability systems—educators want to know how to raise achievement and efficiency. But without causal confidence, their efforts may be on shaky scientific ground.

Given this need for knowledge about what works, the Laboratory for Student Success, the mid-Atlantic Regional Educational Laboratory at Temple University, and the American Psychological Association convened a national invitational conference, "The Scientific Basis of Educational Productivity," on May 13–14, 2004, in Arlington, Virginia, near Washington, DC. This conference was founded on the idea that education research can—and should be—rigorous to contribute substantially to education reform and the improvement of American students' achievement. The commissioned conference papers—written by nationally recognized experts and summarized in this issue of *The LSS Review*—exhibit a variety of scientific approaches to research, emphasizing the special credibility of multiple methods and multiple studies converging on policy- and practice-relevant results.

Assessing the Advantages of Various Research Designs

Experiments are not foolproof in determining causality in education any more than they are in other disciplines: Students in an experimental group may not have been given the full treatment with complete fidelity, the outcome measures may be insensitive, or a Hawthorne effect of being in a special group may elicit greater motivation among experimental students. In well-designed experiments, however, such threats to experimental validity can be enumerated, explicated, and taken into consideration.

Randomized control group experimentation is one of several ways of seeking causal confidence. Single-subject studies, for example, obtain many observations over time while using random, on-again, off-again treatment and control conditions to a single subject.

In some cases, however, randomization may not be feasible. Two of the most currently debated policies in education—school choice and accountability for results—are of keen interest to policymakers, yet neither condition can easily be randomly assigned to schools or school

(continued)

The LSS Review is produced by the **Laboratory for Student Success** (***LSS***) at Temple University Center for Research in Human Development and Education. ***LSS*** is one of ten regional educational laboratories funded by the Institute of Education Sciences of the U.S. Department of Education and seeks to provide research-based information to significantly improve the educational capacity of the mid-Atlantic region, including Delaware, Maryland, New Jersey, Pennsylvania, Washington, DC, and the nation. Inquiries about the work of LSS should be sent to LSS, Outreach and Dissemination Unit, Temple University, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122-6091, lss@temple.edu, or visit the website at http://www.temple.edu/LSS.

districts. Large-scale longitudinal surveys and multivariate statistical methods to control for plausible alternate causes of learning can be used to estimate the effects of such conditions in various contexts. Confidence in all such methods and their results grows when many studies and types of studies yield converging evidence suggesting the same policies and practical implications. Observations and case studies may shed light on how the causal effects operate in specific settings.

Models in Educational Productivity

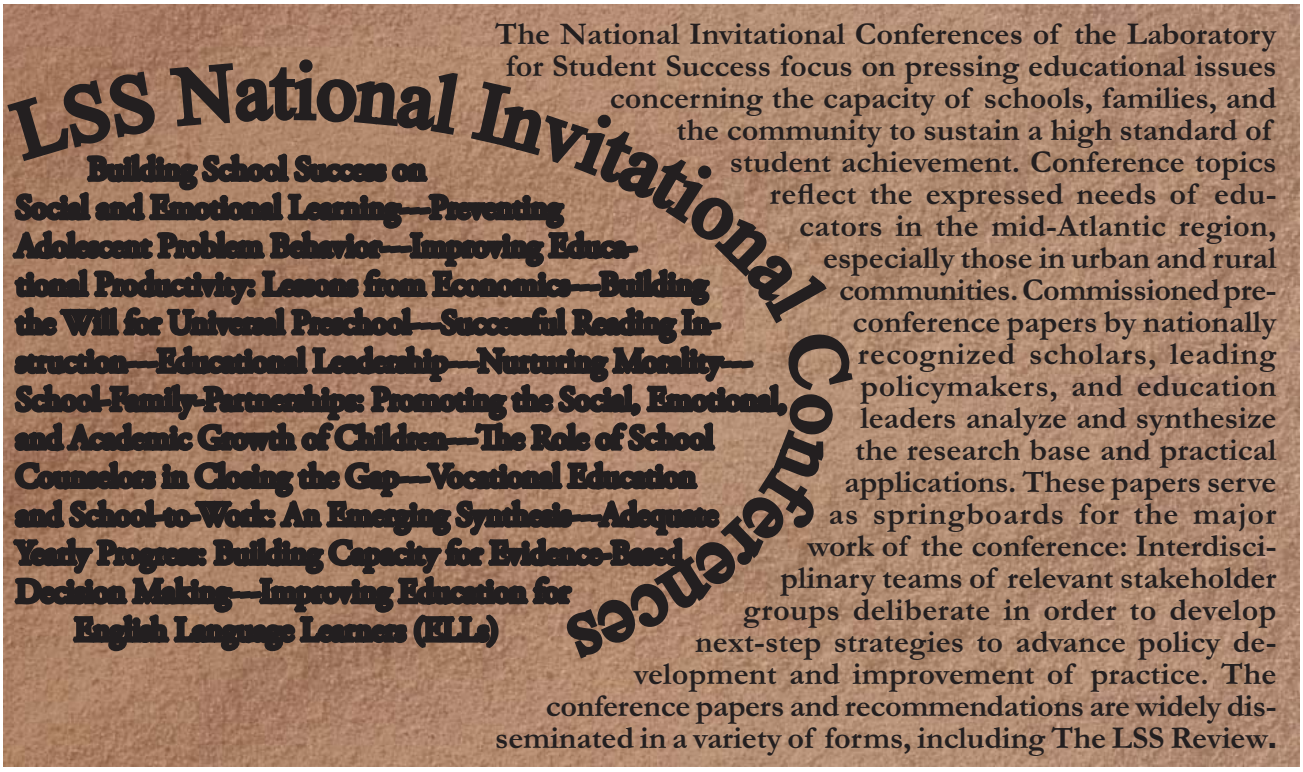
Models should also serve as both the source and product of high-quality research, but they have disadvantages as well. When a model or theoretical framework is applied to school contexts, the central concepts may be misinterpreted or may be poorly implemented. Thoughtful model builders, theoreticians, and educators, nonetheless, can test their ideas in schools. They can work with scientific and educational colleagues to be sure the ideas are well implemented for rigorous testing. They can also draw upon scholarly literature, personal observations, and collegial conversations for support or refutation of ideas before, during, and after data collection.

After an effective program has been devised and extensively tested, policymakers should be informed about it and consider the applicability of the program in professional settings. Considerations other than the effects on learning may weigh heavily. Is the program too costly, too difficult to implement, or not in keeping with their philosophy and values? Even if a program could meet all such criteria, it may not be introduced appropriately. Busy policymakers may not even know about it or the research that supports it. For these reasons, educators too often are left to choose programs based on developer claims, fads, and school or regional traditions. Obviously, more rigorous research and a better means of making the findings known are of high priority.

Content Overview

The articles in this issue of *The LSS Review* may be divided into three groups. The first group reviews selected methods of research. The second explores the development of three models and theories that focus on educational productivity, models relevant to improving students' academic and life skills. Each model has been tested extensively, and each was derived from previous studies that neglected to explain sufficiently how to maximize student potential. The third group of papers describes actual and prospective applications of scientific methods to real problems in education.

The final paper includes recommendations for policy and practice derived from the conference papers and the face-to-face deliberations conducted at the Washington-area conference. The main work of the conference took place in small groups, each representing important scholars and stakeholders in the education community. The task of each group—basing their work on the conference papers, discussion, and their own research and experience—was to develop next steps for applying scientific methods to questions of educational policy and practice and in deriving valid implications from extant and future research. Reported briefly and orally at the end of the conference, the synthesized recommendations serve as the basis for the last paper in this issue of *The LSS Review*.



Experimental and Quasi-Experimental Research Designs

Susan J. Paik, Claremont Graduate University

For decades, poorly designed studies yielded little improvement in student achievement, and the United States continues to rank nearly last in international achievement surveys. Critics have long said that educational research should return to random assignment and control group experiments, and the No Child Left Behind Act of 2001 mandates that educational research be rigorous and "scientifically based."

The methodologies of other disciplines now influence education research, and educational policy and reforms—aiming to improve outcomes—are increasingly based on proven, effective practices. This paper describes three research designs and the hallmarks of good research that help establish such practices.

Types of Experimental Designs

All experimental designs have at least one independent variable and one dependent variable. At least one experimental group receives treatment, but such designs may not have a control group that does not receive treatment. The independent variable is the treatment, and the dependent variable is the outcome. The purpose is to determine if the independent variable causes any changes in the dependent variable.

Pre-experimental Designs

Pre-experimental designs are characterized by a single treated group and no control group; they are susceptible to many threats to validity. *One-shot case studies* measure posttest results without regard to other groups, making them of little scientific value. *One-group pretest–posttest designs* are also flawed because changes may still be attributed to factors outside the treatment despite pretest and posttest comparisons. In a *static-group comparison or posttest-only design with nonequivalent groups*, the experimental group is compared with a comparable group and tested after treatment. Subjects are not randomly assigned. The two groups formed from previously existing groups are intact. However, there is no way to ensure that the two groups started at the same level, negating the value of comparison.

Quasi-experimental Designs

Quasi-experiments are characterized by non-randomly assigning participants to experimental and control groups that may have initially differed. Validity issues are present because participants are not randomly assigned. Quasi-experiments may be the best option when true experiments are not possible. Quasi-experimental designs include both single-group and multiple-group experiments. In the former, the *time series design* involves a series of periodic measurements of a single group before and after treatment. This design, however, lacks control over external influences. Control and experimental participants (single subject or group) are the same individuals in the *equivalent time samples design*, which compares the results of treatment and nontreatment episodes. This recurrent design, however, cannot definitively determine whether the treatment or some unknown is causing the desired effect.

Among multiple-group experiments, the *nonequivalent control group design* uses pretests and posttests for the control and experimental groups. Treatment is not randomly assigned. If it can be assumed that the pretest controls for the group differences, except the treatment, this design comes close to true experimentation's rigor. This design may be one of the most feasible designs in natural settings. More powerful is the *multiple time series design*. It is similar to the single-group time series design except two nonrandomly assigned groups are measured and treatment is randomly applied to one group, followed by postmeasurements.

Experimental Designs

True experiments are characterized by random assignment to experimental and control groups. Compared with previous designs, experiments can better assess a cause-and-effect relationship. True experiments have fewer threats to internal validity than other designs because causal efficacy can be attributed to either the treatment or random differences. However, it is difficult to assign students, teachers, and schools randomly.

In the *pretest–posttest control group design*, participants are randomly assigned to treatment or control groups. Both groups are pretested, and one group receives treatment. Then both groups are posttested, and the scores are compared. In the *posttest-only control group design*, subjects are randomly assigned to two groups. Neither group is pretested. The treatment is given to one group, and the control and experimental groups are posttested. Random assignment reduces possible differences between groups.

Subjects are randomly assigned to four groups in the *Solomon four-group design*. Two groups receive pretests; two do not. One of the two pretested groups receives treatment, and one of the two non-pretested groups receives treatment. All groups receive a posttest. This design may reduce or eliminate the pretest's effect on the outcome, making findings more generalizable.

Hallmarks of Good Research

Random Assignment and Control Groups

Random assignment means that participants have equal chances of being in control or experimental groups. Both groups are considered "equal" in everything but the treatment. Outcome differences are explained by either the treatment or random errors. Although randomization is the ideal, it is difficult and expensive to achieve.

(continued)

Statistical Power and Significance

In lieu of random assignment, *matching* characteristics of participants or schools may provide some confidence in results if experimental and control groups are similar. Matching is inferior to random assignment because causal inferences assume that participants are matched on all relevant factors. It still bears risks of some incomparability, especially if participants drop out of the study.

Experiments and quasi-experiments may be strengthened by using *statistical controls* for groups' initial conditions. These controls usually increase a design's power to detect effects but may not control for all initial variations among groups. *Increasing the sample size* can also strengthen an experiment; the larger the sample size, the greater the power to detect effects—provided the pool is representative of the population. A *power analysis* may be used to identify the appropriate sample sizes. Statistically determined *effect sizes* show how much better (or worse) the experimental groups performed. Effect sizes enable rough comparisons of effects despite use of different tests in more than one study.

Valid and Reliable Measures

Valid and reliable measures are essential to avoid bias. *Nationally standardized tests* base results on national random samples in terms of percentiles or grade equivalents. If these test scores are unavailable, *calibrated developer tests* may be used. However, comparing scores from these two types of tests can be difficult and controversial. When comparing standardized test results, *value-added gains*—the gains from one testing to another—are better than *status scores* (*achievement test score*s), which may be largely determined by students' social advantages or disadvantages.

Internal Validity

Procedures, treatments, or experiences may threaten the validity to draw causal inferences from results. Does the research design rule out any rival explanations? Some threats to internal validity are as follows:

- *History* is associated with unanticipated events on the participants after the pretest.
- *Maturation* includes biological and psychological processes between the pretest and posttest.
- *Instrument or instrument decay* includes changes in measurement, including changes in standards of classroom observers or test scorers over time.
- *Testing* is the effect of a pretest on a posttest. Retesting may affect the scores of a second test despite treatment.
- *Statistical regression* occurs when students who score low do better on retesting and those who score high do worse. Both kinds of scores regress closer to the mean.
- *Mortality* or *attrition* is the differential loss of subjects that may influence a group mean.
- *Selection bias* can occur when subjects are not randomly assigned to groups; nonequivalent groups affect the dependent variable.
- *Treatment interactions* include effects on the treatment by another treatment, situation, or participant group.
- *Selection–maturation interaction* occurs when maturation is not consistent across groups because of selection bias.

External Validity

An externally valid experiment is one that applies one design to different participants in different locations and, using the same measures, reproduces similar results. Can the same findings be expected with other people, settings, and times? Some threats to external validity include:

- *Obtrusiveness and reactivity*: Participants react to obtrusive treatment, affecting results. In the *Hawthorne effect*, subjects may improve their performance when they think they are receiving special attention. *Hypothesis guessing* occurs when participants respond as expected. *Compensatory rivalry* may occur in educational research when control teachers work harder to make up for any loss of benefit for their students. The *novelty effect* occurs when the treatment is new and interesting, with participants initially responsive and then later losing their zeal.
- *Researcher expectancy effect or Pygmalion effect*: Researchers may inadvertently tip the scales toward a desired effect. Researcher expectancy effects also occur when researchers know who is receiving the treatment. Therefore, observers, treatment administrators, and subjects should be "blind" to the experiment's features.

Analytic Validity

Findings may be threatened by analytic validity problems. Discussions of external validity neglect some of these analytic threats except for power, error rate, and reliability. Some of the often-neglected threats are as follows:

- *Leveling* occurs when continuous measurements are grouped into arbitrary levels, such as high, middle, and low, thus losing the original precision of the variables. Regression analysis is preferable to analysis of variance because it does not require leveling.
- *Outliers* are mismeasures or mistaken data that may produce interaction, reversals, curvature, and abnormal residuals.
- *Colinearity* occurs when independent variables are correlated or co-occur, making separation of their effects difficult.

Conclusions and Recommendations

We need well-designed studies to help interpret the causes behind learning. To improve our understanding of learning, studies' internal and external validity must be strengthened. Although random assignment to experimental and control groups is the ideal of scientifically based research, it may not always be feasible because of cost, time, or ethics. Quasi-experiments may not meet this ideal but may yield valid conclusions when well designed. The best evidence may be found in the consistency of a multiplicity of study results, and meta-analyses of many well-designed experiments can provide confident causal conclusions.

Scientific Formative Evaluation: The Role of Individual Learners in Generating and Predicting Successful Educational Outcomes

T.V. Joe Layng, Greg Stikeleather, and Janet S. Twyman, Headsprout

The effort to bring scientific verification to the development and testing of educational products and practices has earnestly begun, as required under 2001's No Child Left Behind Act. Products used for teaching reading are the first targeted for improvement. Other products and practices will follow, particularly if the initial effort successfully impacts children's reading performance. This paper examines the role of scientific evaluation in enhancing instructional productivity through the application of rigorous scientific evaluation, both during development of instructional programs and in their larger scale validation.

Definitions

Early reading programs are often self-described as *research based*, but their use of the term reveals a remarkably wide range of meanings. For some instructional programs, it means merely that they claim to contain elements that research suggests are effective. For other programs, it indicates that pretest and posttest or simple comparison studies have provided some evidence of effectiveness. For still others, it describes some form of scientifically controlled study, often involving randomized, control groups. This imprecision in the term's use is compounded by its failure to distinguish between a program's *scientific development* and the *scientific evaluation of outcomes* after development.

This latter use of *research based* might more properly be considered *research filtered:* Regardless of how a program was designed, it is measured against an alternative form of instruction or sometimes no instruction at all. This sense of *research based* refers to an emphasis on *summative* evaluation. In the research-filtered approach, the program itself does not have to be scientifically designed or based on research.

Another use of the term *research based* might be more properly considered as *research guided*, which refers to a program of instruction that has been scientifically designed and tested during its development, or at least its design is guided by previous research results. This sense of *research based* refers to an emphasis on *formative* evaluation. In the research-guided approach, formative evaluation is intertwined into the instructional design protocols and, at its most rigorous, influences program development through iterations of testing, revising, and retesting.

Levels of Verification

Both formative and summative evaluation may evidence varying degrees of verification based on their commitment to a scientific approach, ranging from experiential, to evidence based, to scientific. In the more rigorous forms of formative evaluation (also referred to as *developmental testing*), data are continuously collected and analyzed during program development to provide an ongoing *experimentally controlled* research base for ensuring effectiveness with *individuals*. In the more thorough forms of summative evaluation, data from randomized experimental and control groups are collected and analyzed to provide a *statistically controlled* research base for determining program effectiveness with *groups*.

In the least thorough forms of formative and summative evaluation, emphasis is placed on philosophy, point of view, and anecdotal evidence. Little attention is paid to direct measurement of instructional effect or to the determination of functional relations among variables. Both forms of evaluation also have middle grounds that include attempts to use some form of empirical evidence, influence program development (for formative evaluation), and make judgments about outcomes (for summative evaluation).

Implications for Research-Based Instruction

Programs that evolve from a rigorous formative evaluation process can predict *individual* outcomes across all summative evaluation levels of rigor, just as programs tested under the most rigorous form of summative evaluation can predict *group* outcomes across all formative evaluation levels. Both should be considered to have equal predictive power. Both formative and summative evaluation are important and may be combined to provide useful information on individual performance and group averages.

At its most rigorous, formative evaluation requires a careful control analysis design to ensure that each part of the program works alone or together with other parts to produce a predictable outcome. Accordingly, such formative evaluation lends itself most readily to single-subject research designs in which participants respond over long time periods while variables are experimentally changed and controlled. In these designs, variance is controlled through direct procedural or experimental intervention, rather than through group assignment.

Whereas group designs are readily known and accepted as providing scientific evidence for program effectiveness, single-subject designs are not so well known. Although both group and single-subject designs are descended from highly successful scientific traditions

(continued)

and may provide equally rigorous results, single-subject designs are less well understood.

Single-Subject Control–Analysis Evaluation

Single-subject designs are most valuable when the questions addressed concern how program components working alone or together affect an individual's performance. These designs provide predictions on how *individuals*, rather than groups, using the program will perform compared with a standard. In group experimental designs, statistical controls and analysis are used to account for variance, often using randomized or matched control groups. But in single-subject experimental designs, procedural change attempts to directly control variance.

Although sharing the goal of predicting program outcomes with summative evaluation, the procedural control–analysis designs, which typify formative evaluation, differ from summative evaluation and statistical control designs in another important aspect. In single-subject research designs, the essential question is whether experimental control is maintained over the learner's behavior as response criteria are systematically changed. Also, after such control can be demonstrated for an individual, the research questions whether that control can be replicated for other individuals across different settings.

In such systematic replication, the occurrence of increased variance in responding, both within a learner's individual performance and between the performance of different learners, allows for the examination of the program elements and sequence in which the variance occurred and the modification of (or the design of new) procedures to reduce or control the variance found in meeting the mastery criteria. Systematic replication with new individuals provides increased confidence that the same procedures will provide similar outcomes for other individuals. Each new learner can be considered an experimental replication.

Emphasis on the Individual

Scientists and engineers who design and build complex systems rely on rigorous formative evaluation. Testing helps the designers determine if the designs are working and make modifications to improve stability and reliability. Each testing is considered a replication; the more conditions encountered, the more systematic the replication. One product is not constructed and then compared with other products to determine if it works better than differently built products comprising a control group. Rather, each revision based on the testing is retested until the component meets a quality standard. Only after rigorous testing of the components, both separately and together, is the final question asked: "Does it work?"

Rigorous formative evaluation may have a similar effect on teaching reading and other instructional program development. By ensuring that each component meets a specified quality standard, educational researchers should be able to design and build instructional programs that have the same high likelihood of success as does the building of industrial products. Rigorous "single-subject" test–revise–retest cycles provide great confidence that all products built in accord with the design and development process will work correctly without the need for tests comparing groups of products. A similar approach to educational program development may provide comparable confidence.

When rigorous formative evaluation is not possible, the only recourse is summative evaluation. Here, statistical (rather than direct experimental) investigation is used to evaluate the efficacy of the procedures or treatment being developed. But exclusively relying on summative evaluation protocols may not necessarily be the most informative approach for assessing instructional programs.

A rigorous research-guided formative evaluation applied to designing and building instructional programs tells more than that the "mean" experimental child performs better than the "mean" control group child. It tells how the program components work separately and together and whether the components are effective with each individual. By setting formative evaluation criteria high, researchers may be able to ensure that nearly all children who use a program so developed succeed.

Such an assertion is quite different from merely stating that the experimental group performed significantly better than did the control group, because, with a large enough number of participants, small absolute differences between groups can produce highly significant results. Rather than producing instructional programs that work only, on average, better than other programs—even if that outcome has been "scientifically" determined—it may be better scientifically (or socially) to produce programs that are the products of rigorous formative evaluation and that must work, therefore, with each individual.

Large-scale Formative Evaluation

A scientific formative evaluation can also play an important role in improving educational productivity at the district and school level. Although rare, some school districts are working diligently to put in place sophisticated data-gathering instruments and frequent assessment so that educational practices at the classroom level can be tested, revised, and retested until the practices are successful as measured against a standard. As a result, in these cases in which careful formative evaluation practices have been used over time, entire school districts have begun to make progress in closing the achievement gap between majority and minority students—often with both goups achieving at much higher levels.

Blending Experimental and Descriptive Research: The Case of Educating Reading Teachers

Elizabeth S. Pang, Ministry of Education, Singapore, and Michael L. Kamil, Stanford University

The importance of teacher preparation and professional development in educational research has generated considerable professional and public interest as a result of school reform efforts that are predicated on having highly qualified and committed teachers. Teacher preparation and quality have come under close scrutiny because of the U.S. Department of Education's (USDOE) 2002 report, which stated, "There is little evidence that education school coursework leads to improved student achievement."

In view of the paucity of evidence reported and in view of the USDOE's call for new standards in teacher education, we reexamine the state of research in literacy teacher education, both preservice and inservice. We draw upon a review of experimental and quasi-experimental research published by the National Reading Panel (NRP) in 2000 and update and expand the coverage of that review by examining studies that use correlational, descriptive, or other methodologies. We assert that in order to synthesize findings and make recommendations that affect policy and improve reading teacher education, researchers need to examine critically the methodologies of studies in order to come to a holistic understanding of what works and why.

Teacher Education Research

Key questions facing researchers and policymakers are whether teacher education programs are effective in changing teachers' knowledge and practices and whether such changes, if they occur, increase student learning. Answers to these questions will help determine the characteristics of effective programs for reading teachers.

Because there are multiple layers of causal relationships, encompassing teacher educators to students and including materials and environment, researchers typically focus on a few processes of teaching and learning at a time, usually using a specific method to answer the research question. Research on instructional variables, for example, generally examines the interactions between teachers and students in particular contexts of learning.

The Database

We identified 306 studies published between 1961 and 2001 and divided these into experimental/quasi-experimental and non-experimental studies. We coded the studies, analyzed the overall trends, and closely examined particular themes and issues emerging from groups of studies. Finally, we compared the findings of the experimental/quasi-experimental and descriptive research to derive principles and practices that promote both teacher learning and student achievement.

Experimental vs. Non-experimental Studies

The number of non-experimental studies far exceeded those of experimental and quasi-experimental studies. There were also far more studies of preservice teachers than of inservice teachers. Experimental studies provide causal evidence of teacher improvement and sometimes of concurrent student achievement, and non-experimental studies use a variety of approaches and methodologies, providing multiple perspectives and rich contextual descriptions of teacher learning. Correlational data suggest that certain aspects of teacher quality characteristics such as certification status and degree in the field to be taught are positively correlated with student outcomes. However, this does not tell us if these characteristics ultimately lead to better student achievement.

Findings of Experimental Research

The NRP report highlights the need to measure both teacher change and student outcomes to demonstrate the effectiveness of teacher education. However, these conditions are often difficult to achieve for preservice education. For example, the NRP reports, as might be expected, that 10 of 11 of the preservice experimental studies revealed improvements in the knowledge of prospective teachers, but it is unknown whether their new learning impacts classroom practice and student learning. A longitudinal study would have to follow preservice teachers into their first year of teaching and beyond. Given the differences between sites where teachers from the same programs teach, the power of such a study would be relatively slight, so few of these studies have been done.

The problems are not as severe for studying inservice education because these sites are identifiable and accessible. However, only 11 of 21 inservice experimental studies reported both teacher and student outcomes. The majority of studies that measured either teacher or student outcomes showed significant or modest improvements in either teacher knowledge or student achievement. Those that measured both provide clear evidence that inservice teachers do learn from professional development programs focusing on specific types of reading instruction and that students of those teachers benefited from improved teaching.

Findings of Non-experimental Research

Non-experimental designs predominate in preservice studies because of researchers' interest in relating teachers' learning processes, both individually and collectively, to prescribed coursework, field experience, or combina-

(continued)

tions of these. In general, these non-experimental studies affirm the importance of providing field-based experiences in conjunction with coursework in order to help teachers connect theory and practice. The majority also report favorably on preservice teacher change, but it is uncertain if this change leads to application, though some research suggests that the use of preservice training becomes increasingly evident in the first two years of teaching. Still uncertain is the effect on student learning. Few of these studies measure or report student outcomes.

In concert with trends in preprofessional preparation of teachers, substantial numbers of non-experimental studies have focused on the variously conceived practice of reflection to examine the process of change in prospective teachers' beliefs and attitudes in relation to a host of instructional issues. Similarly, the importance of technology has stimulated numerous non-experimental studies of the impact of new technologies on literacy teacher education largely ignored by experimental research—multimedia, hypermedia, and computer-mediated communication. Non-experimental studies have also been instrumental in foregrounding the under-researched issue of teaching reading to culturally diverse learners.

Non-experimental studies of inservice professional development, as with experimental studies, focused on more specific instructional methods and issues compared with preservice studies. Conceptual tools buttressed with practical strategies prove to be the most influential, and conferencing with mentors and supervisors is also important.

Future Directions

Experimental research provides evidence of teacher change and its effect on student achievement. To guide change more effectively, we must also understand more deeply teachers' attitudes, beliefs, and conceptualizations of literacy and the changes they undergo while studying practices and outcomes; knowing about the beliefs and attitudes of teachers is important because it indexes a source of teacher behaviors. In one study, for example, correlational analyses indicated that teachers' philosophical acceptance predicted their use of instructional methods. Improving and non-improving teachers were different in their self-efficacy and willingness to experiment. Because non-experimental studies ask questions different from those asked by experimental studies—focusing on the processes of change and reflection—both kinds of studies are needed. The findings of the non-experimental studies of teacher change do not contradict those of the experimental research, but they need to be designed and reported better to facilitate parallel or follow-up studies. Furthermore, more longitudinal studies that track teachers through their initial years of teaching and studies investigating diversity need to be rigorously pursued.

What Does This Analysis Reveal About Research?

One of the key assumptions held about teacher education and professional development is that, if it is effective, it should produce "better" instruction (changes in teacher behaviors) and "better" reading by students (higher achievement). However, this assumption does not drive much of the research. Only some experimental studies compared groups as well as outcome measures, but both are needed for establishing links between interventions and performance. Improvements in research conceptualization and design would allow such analyses to be conducted—analyses that are key for policy work, for example, in establishing the relative costs of raising reading achievement through different professional development programs.

Improvements in methodology and reporting could also lead to a more integrated and holistic understanding of teaching reading. Some of the imbalance between the numbers of experimental and non-experimental studies and between preservice and inservice studies can be accounted for by costs, by the questions being asked, or, regrettably, by assuming that researchers chose a methodology and then found a problem to study. This latter tactic may become less attractive because of the current national policy, which has adopted as its exemplary standard the experimental research design. It must be acknowledged that non-experimental methodologies may be preferable for certain problems. Integration of knowledge would be facilitated, at least, by authors making their assumptions explicit when reporting studies and by journal editors requiring explicit statements of the relationship between questions, methodology, and data. Finally, we observe that researchers rarely cite relevant research from paradigms other than their own. But much can be gained by having authors broaden their view to include research from different methodologies. Blending data from research conducted using different methodologies has the potential to enrich the knowledge base.

* * *

The Laboratory for Student Success (*LSS*) is one of the nation's 10 regional educational laboratories funded by the Institute of Education Sciences (IES) of the U.S. Department of Education to revitalize and reform educational practices in the service of the educational success of the nation's children and youth.

The primary mission of ***LSS*** is to bring about lasting improvements in the learning of the mid-Atlantic region's increasingly diverse student population. ***LSS*** seeks to establish a system of research, development, and dissemination that connects schools, parents, community agencies, professional groups, and higher education institutions in order to transform low-performing schools into high-performing learning communities.

* * *

The Enhancement of Critical Thinking

Diane F. Halpern, Claremont McKenna College

Among its stated aims, the Goals 2000 Educate American Act, enacted in 1994, promised to increase the proportion of college graduates demonstrating an advanced ability to think critically. That the government made this commitment and set a date for achieving results belatedly recognized the need to teach for critical thinking. Regrettably, this goal was neither funded nor implemented on the same national scale as others stipulated in the act. Nevertheless, many colleges voluntarily instituted required or optional critical thinking courses. Courses with a similar purpose were subsequently adapted for secondary schools, but too much of what is learned in U.S. schools is closer to rote learning than it is to learning in ways that promote critical thinking.

Much literature is available on programs to teach critical thinking, and a substantial amount of evidence indicates critical thinking can be taught and learned, especially when instruction is specifically designed to encourage transfer of skills. Nevertheless, the types of studies required to confirm with certitude the efficacy of teaching critical thinking present practical and methodological problems.

Critical Thinking

Most definitions of critical thinking refer to the mental processes of reasoning logically, making judgments, questioning, and reflecting on the process itself. I define the term in the following manner: Critical thinking is the use of those cognitive skills or strategies that increase the probability of a desirable outcome. It is thinking that is purposeful, reasoned, and goal directed—the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions. The use of the word "critical" in "critical thinking" is used in the sense of judgment. Critical thinking includes evaluating the quality and outcome of the thinking process.

A Skills Approach

Critical thinking instruction that is skill based has specific educational objectives—and thus is easier to assess and communicate to students and other stakeholders—and provides a framework to focus classroom lessons. Some examples of thinking skills, applicable in a wide range of situations, are understanding how cause is determined, recognizing and criticizing assumptions, analyzing means–goals relationships, supplying reasoned support for conclusions, assessing probability, incorporating isolated data into a wider framework, and using analogies to solve problems.

Transcontextual Transfer

Thinking skills can be taught and transferred to other topics. Transfer is the spontaneous use of a skill in a context different from the one in which it was learned and is the goal of critical thinking instruction. The failure to transfer a skill maybe attributed to inadequate learning of the skill or teaching that does not encourage transfer. When skills are taught for transfer—with multiple examples across different domains of knowledge, uncued but with corrective feedback—they do transfer. Such teaching should include direct instruction with review, teacher modeling, guided and spaced practice, and independent application.

Assessment as an Operational Definition

The assessment of an intervention is almost as important as the intervention itself. Assessment is tied to issues of definition, research design, and essential debates over whether it is possible to improve thinking. "Off-the-shelf" critical thinking assessments are available, but they rarely match what is taught in critical thinking courses, and many of the tests have very poor psychometric properties. When the measurement is bad, it is easy to see why we have not gotten strong results with critical thinking instruction, but the measurement issues in critical thinking are not insurmountable.

A Better Measure

The need for providing information about the status of critical thinking skills is relatively uncontroversial. There is currently little information to inform decision makers concerned with improving thinking skills. The controversies arise over questions of whether the information can be provided in a way that is meaningful, valid, fair, and cost effective. If the assessment is not well done, the results will be costly. A good measure of critical thinking would be based on clearly defined skills assessed in realistic scenarios that could apply to a wide range of ethnic and socioeconomic groups. The skills selected must ones used in most cultures.

The Sequential Question

The Critical Thinking Assessment About Everyday Events uses realistic examples with an open-ended response format, allowing participants to demonstrate spontaneous use of skills. Participants are then probed for alternatives in forced-choice questions, demonstrating their understanding of concepts and showing if they are able to use skills when prompted.

A good critical thinking question with several sequential parts allows for different types of information about participants with a minimal number of questions. Open-ended parts test "free recall" because they place few

(continued)

restraints on responses. Multiple-choice parts show if respondents are able to recognize skills presented in a list, a measure of "recognition memory." These two types of recall use different cognitive processes. Lower scores are expected on free recall tests because they require a search through memory plus a verification of answers; recognition requires only the verification stage.

Tests presented on computers provide reaction time data, which help provide information about the microcomponents of the underlying cognitive processes. Reaction times permit a much more fine-grained analysis of mental events than other commonly used dependent measures.

Cognitive psychologists can now provide sufficient knowledge of how people think, learn, and remember. People retain information best when they generate information from memory, space practice over increasing time intervals, remain active, receive informational and useful feedback, and use visuospatial and verbal formats.

Thorny Conceptual Issues

The literature on teaching thinking skills is huge but difficult to summarize statistically because of the variety of instructional strategies—team teaching, learning hierarchies, tutoring, questioning, and concept mapping, to name a few—that have been investigated. Random assignment field trials may be proposed as a way to confirm the efficacy of teaching critical thinking, but this supposition is based on imperfect analogy between education and medicine: We do not improve thinking the same way we prevent polio. Furthermore, the many criticisms of null hypothesis testing cannot be "fixed" by randomly assigning participants to conditions. Alternatively, meta-analyses might allow for information across studies to be considered along with a single estimate of their effect size, but such meta-analyses raise the issue of how multiple studies with large effect sizes with a matched control group should be weighed against a single experiment with random assignment of subjects and a smaller effect size. One synthesis of studies of thinking skills programs computed an overall effect size of d = 1.17 from 45 separate effect sizes. With an effect size over one standard deviation across studies with diverse subjects and settings, as was the case in this synthesis, do we need large random assignment field trials before we can decide that these interventions work to improve thinking skills? Furthermore, would informed parents allow their children to be in the control or non-treatment group in a randomized trial? The "correct" response is that without random assignment, it is not possible to know if the intervention actually worked. However, a large effect size summarized over a large number of diverse studies from many different participants and contexts also provides good evidence, even if it is not strictly causal.

Avoiding Design Flaws

Even though they are inherently flawed, we absolutely need large-scale random assignment studies, but we need to be mindful of their limitations and not blindly accept conclusions as "the answer" to questions. We can also use meta-analyses that indicate effect sizes and other types of converging evidence. The complexities of real children in real learning environments do not easily lend themselves to the manipulation of single variables under controlled conditions, but these sorts of studies need to be funded and encouraged or they will not happen because of the necessary expenses and need for replication, fidelity, and collaboration.

Strong Causal Evidence

One large-scale, double-blind, random assignment experiment of a thinking skills intervention showed that targeted thinking skills were transferred and used appropriately with novel topics. Students who received the thinking skills instruction showed greater gains than control group students on tests of general aptitude, problem solving, decision making, reasoning, creative thinking, and language. Improvements in thinking are possible when instruction is designed for this purpose.

Large-scale studies of the type described above are expensive and need governmental or foundational support, but such studies are needed so that results can be replicated across sites and so researchers can establish necessary controls to determine the effect sizes. Conclusions from studies with poor controls suggest that low-achieving students make the greatest gains, perhaps because they have the greater possible latitude for additional cognitive gains, but experiments are needed to verify this.

Conclusions

Students can think better as a result of instruction, but we lack the strongest causal data with longitudinal follow-up. More randomized field trials are needed. These studies are expensive and difficult to coordinate but are worth the investment. Educated adults need to be able to judge the credibility of information, recognize and defend against propaganda, reason effectively, use evidence in decision making, and identify problems and find solutions if they are to benefit from the wealth of available information. Doing all this may be the best return on investment we make as a nation.

Improving Educational Productivity: An Assessment of Extant Research

Herbert J. Walberg, University of Illinois at Chicago

The purpose of this report is to synthesize (a) meta-analyses (statistical analyses of results of many studies) of control group research and (b) large-scale surveys that reveal the causes of academic achievement. Although economic, sociological, and political factors affect learning, their influence is indirect. Learning is fundamentally a psychological process; student motivation, instruction, and other psychological factors are the well-established, consistent, and proximal causes of learning.

An early synthesis of 2,575 study comparisons suggested nine factors, in three areas, that are the chief psychological causes of academic achievement (and, more broadly, school-related cognitive, affective, and behavioral learning):

A. Student Aptitude

1. Ability or preferably prior achievement
2. Development as indexed by chronological age or stage of maturation
3. Motivation or self-concept as indicated by personality tests or the student's willingness to persevere intensively on learning tasks

B. Instruction

4. Amount of time students engage in learning
5. Quality of the instructional experience, including method (psychological) and curricular (content) aspects

C. Psychological Environments

6. Morale or student perception of classroom social group
7. Home environment or "curriculum of the home"
8. Peer group outside school
9. Minimal leisure-time mass media exposure, particularly television

Subsequent syntheses have shown results consistent with the original findings. Each of the first five factors—prior achievement, development, motivation, and the quantity and quality of instruction—seems necessary for learning in school. Without at least a small amount of each, the student may learn little. Large amounts of instruction and high degrees of ability, for example, may count for little if students are unmotivated or instruction is unsuitable. Each of the first five factors appears necessary but insufficient by itself for effective learning.

Time is a particularly pervasive constraint, since U.S. students have the shortest school year among countries of the industrialized world and generally do far less homework than students from other countries. Until recently, with the advent of summer and after-school programs, time remained neglected among school reforms. The positive effect of time is perhaps the most consistent of all causes of learning.

In addition to time, intensity is also very important: Illogical or unsuitable instruction or student inattentiveness may mean that little is accomplished, notwithstanding much study time. Other psychological conditions also have a causal bearing on learning.

The four psychological environments listed above can expand and enhance learning time. Classroom morale is measured by obtaining student ratings of their perceptions of the classroom group. Good morale means that the class members like one another, that they have a clear idea of the classroom goals, and that the lessons are matched to their abilities and interests; in general, morale is the degree to which students are concentrating on learning rather diverting their energies because of unconstructive social climates.

Peer groups outside school and stimulating home environments can help by expanding learning time and enhancing its efficiency; students can both learn in these environments and become more able to learn in formal schooling. The last factor, mass media, particularly television, can displace homework, leisure reading, and other academically stimulating activities; and it may dull the student's keenness for academic work. For instance, some of the average of 20 to 30 hours a week high-school students spend viewing television might usefully be added to the mere 4 or 5 average weekly hours of homework they report.

Three of the nine factors require close attention: quantity and quality of instruction, because educators can alter these factors, and the home environment, because it influences the large amounts of time students spend outside school and because it can be affected by outreach programs. In case studies of poor inner-city Chicago families, the children who succeeded in school had parents who emphasized and supported their children's academic efforts, encouraged them to read, and interceded on their behalf at school. Many statistical studies show that indexes of such parent behaviors predict children's academic achievement much better than socioeconomic status (SES) and poverty. Cooperative efforts by parents and educators to modify alterable academically stimulating conditions in the home have had beneficial effects on learning. In 29 controlled comparisons, 91% of the comparisons favored children in such programs over nonparticipant control groups. Although the average effect was twice that of SES, some programs had effects 10 times as large, and the programs appear to benefit older as well as younger students.

At-Risk Students

Sizable proportions of young children, especially those in poverty, are behind in language and other skills

(continued)

before they begin school, and they are often placed in bilingual and special-education programs for the developmentally challenged, in which they are segregated from other children and make poor progress. The origins of their achievement problems are partly attributable to ineffective programs, but the origins can also often be traced to specific parental behaviors before children begin school that affect children's reading and other language skills, which are keys to achievement in academic subjects.

The growing gap between good and poor readers reflects social class differences. These differences stem from early childhood experience, especially with respect to parent behaviors that motivate children. Studies show that middle-class parents are more likely to hold high expectations for their children's achievement and to be more often engaged with them in promoting it. Differences in vocabulary related to SES are strongly associated with parent behaviors. Higher SES parents spend more minutes per hour interacting with their children and speak to them more frequently.

These patterns are hardly inevitable. In 47 states and the District of Columbia, effective education policies and teaching practices have enabled more than 4,500 high-poverty and high-minority schools (high meaning over 50%) to perform among the top one third of schools in their states and often to outperform predominantly White schools in advantaged communities. These schools educate about 1,280,000 low-income students, about 564,000 Black students, and about 660,000 Latino students (the groups overlap).

How do these schools do it? Their principals tend to report the following features of their schools: extensive use of state/local standards to design curriculum and instruction, assessment of student work, and evaluation of teachers; increased instruction time for reading and mathematics; substantial investment in professional development for teachers focused on instructional practices to help students meet academic standards; comprehensive systems to monitor individual student performance and help struggling students before they fall behind; parental involvement in efforts to get students to meet standards; state or district accountability systems with real consequences for adults in the school; and use of assessments to help guide instruction and resources and as a healthy part of everyday teaching and learning.

The only long-term study of an academically focused, school-related program showed significant long-term effects and cost effectiveness. The Chicago Child–Parent Centers (CPC) provided academic and family support services to children, beginning at age 3. The program emphasized the acquisition of language and prematheматical experiences through teacher-directed, whole-class instruction, small-group activities, and field trips. Parental participation in the program was intensive. Compared with matched control group children, the 989 CPC children in the program showed higher cognitive skills at the beginning and end of kindergarten, and they maintained greater school achievement through the later grades. By age 20, CPC graduates had substantially lower rates of special-education placement and grade retention than the control group, a 29% higher rate of school completion, and a 33% lower rate of juvenile arrest.

Effective, efficient classroom teaching methods can diminish gaps between abilities and raise all students' achievement, yet costly methods and conditions remain prevalent in U.S. schools.

Effective Policies

A Nation at Risk and subsequent reports showed Americans the importance of achievement for national and individual prosperity and welfare. The congressionally commissioned National Assessment of Educational Progress, however, has shown little achievement change since then, which has led to increasingly substantial reforms. At the school, district, and state levels, some policies have shown positive learning effects, including accountability, incentives, external examinations, and small schools and small districts.

ACCOUNTABILITY

In 1989, the National Governors' Association "Education Summit," with then President George Bush and business leaders, gave impetus to business-style accountability for schools. "Systemic reform," as recommended by summiteers, meant aligning the chief parts of school systems with one another, specifically fitting state tests and curricula with state goals or standards and making exam results widely known. State policymakers set goals, measured progress, and encouraged local school districts and schools to plan and execute effective practices. State officials set high targets for achievement or value-added learning gains, while maintaining more objectivity in evaluating the results than when they determined both goals and means. Without this division of labor, local districts might set easy-to-reach, unmeasurable, or obfuscated goals. Large-scale research on school accountability shows strong public recognition of the need for accountability and corroborates the expected positive learning effect. Policy analysts have begun rating the states for both standards and accountability, which to be most effective, must presumably go together. Good standards are rigorous, clear, written in plain English, communicate what is expected of students, and can be assessed. Good accountability systems are aligned with the standards and include school report cards, ratings of schools, rewards for successful schools, authority to reconstitute failing schools (e.g., by replacing the staff), and the actual exercise of such legislated consequences.

INCENTIVES

Similarly, student incentives, particularly high standards, promote learning. The threat of grade retention, for example, can serve as an incentive for greater effort, although intensive remediation seems necessary. An example is Chicago's Summer Bridge program, which

gave parents and students the choice of grade retention or passing an intensive, focused summer course. Depending on the grade level and subject, grade-equivalent increases in reading and mathematics scores over the short summer session ranged from one half to a full year. The gains were extraordinarily effective, time efficient, and cost effective; and they were sustained in subsequent school years. Tough grading standards and required homework also benefit learning. Requiring high-quality work for a given assigned grade generally raises achievement, particularly for high-achieving students who might not otherwise be sufficiently challenged.

External Examinations

The Cornell economist John Bishop intensively studied effects of curriculum-based external examination effects on learning. He analyzed surveys of the examination effects on learning of the (U.S.) Advanced Placement program, the New York State Regents, and U.S. state and Canadian provincial systems. He also analyzed examination effects on learning in the United States in comparison with effects in Asian and European nations. The examinations have the common elements of being externally composed and geared toward agreed-upon subject matter students are to learn within a nation, state, or province. Often given at the end of related courses, the examinations have substantial positive effects on learning. Made publicly available, the examinations allow citizens, policymakers, educators, parents, and students to assess and compare achievement standings and progress. The largest and most sophisticated international comparative analysis of national achievement yet conducted corroborates Bishop's and related findings. Using data from 39 countries that participated in the Third International Mathematics and Science Study, a Kiel (Germany) Institute of World Economics study found that nations where students learned most employed external, curriculum-based examinations, and policymakers closely monitored the results.

Small Schools and Small Districts

The psychological and economic advantages of small schools and small districts make them more effective and efficient. Of course, after much painful district consolidation and huge capital investments in large school buildings, the clock cannot easily be turned back. But it can be recommended that districts think twice about further consolidation and building ever-larger schools. More radically, legislators have been considering the breakup of large districts such as Los Angeles and New York into completely freestanding units with separate boards and superintendents. Citizens in parts of Los Angeles are pressuring legislators to allow secession. Big urban districts such as Chicago and New York foster "schools within schools" that are attempts to recover the intimacy, accountability, effectiveness, and efficiency of smaller schools of yesteryear, though it remains to be seen if such values can be recaptured in big buildings. Because large districts are less effective and efficient than small districts, special forms of accountability seem necessary in large districts to ensure effective, efficient schools that are satisfying to parents.

Productivity Deterrents

Many prevalent education policies and practices are unproductive. Some take time away from what works consistently and well, and some are costly, disruptive, distracting, and have unanticipated harmful consequences. Their prevalence helps explain why American students fall behind despite high and substantially rising expenditures. For the sake of increasing educational productivity, it is worth considering and avoiding them. They range from applications of pseudoscientific psychology to categorical federal education policies.

Widespread, Unsubstantiated Programs

Some influential education theorists and educators oppose accountability, standards, testing, and the evidence-based learning principles discussed above, most of which comport with what the legislators, public, parents, and students themselves expect from schools. Many in education have been subject to pseudoscientific fads for which there is usually initial enthusiasm but paltry or analogous evidence. At the request of the U.S. Army, for example, the National Academy of Sciences evaluated exotic techniques and "shortcuts" for "super-learning" described in "pop" psychology. Little or no evidence, however, was found for the efficacy of learning during sleep, mental imaging of motor skills, "integration" of left and right brain hemispheres, biofeedback, and such parapsychological techniques as extrasensory perception, mental telepathy, and mind-over-matter exercises. Even so, "brain-based learning" is gathering momentum in education circles.

School-board members and most educators lack education and experience in accountability, evaluation, and methods of psychometrics and statistics that would enable them to choose effective, efficient programs and weed out others. Though these tasks should be central to leaders aiming to measure, evaluate, and improve learning, they are neglected. Consequently, popular programs are often chosen by fad and reputation rather than by a careful review of evidence of their results and costs. Two widespread programs—Reading Recovery and Success for All—illustrate such choices.

Begun in 1976 in New Zealand, Reading Recovery was implemented in 40 states within 8 years. Because Reading Recovery teachers tutor a single student pulled out of regular classes for long periods, Reading Recovery students lose time in regular instruction. The annual per-student cost for the program tutoring alone, moreover, is at least three quarters that of a full program for other students in all subjects all day for the school year. In contrast, phonics, phonological awareness, and repeated oral reading instruction have substantial effects and can be employed cheaply, routinely, and effectively with a whole class.

(continued)

Independent studies of the Success for All (SFA) program have shown no effects. Contrary to SFA claims, average SFA third graders were not up to grade level in Baltimore, where the program originated; by fifth grade, they were 2.5 years behind. Despite the fact that SFA schools were given substantially more funds, materials, and services, independent evaluations showed SFA schools do about the same as control schools.

Federal Categorical Programs

Two very costly federal programs—Title I and special education—also have poor records of promoting achievement. The federal government has spent more than $125 billion on Title I. The program was to have reduced the gap between middle-class students, often Whites in suburbs on the one hand, and on the other, poor students, often African Americans and Hispanics in cities. Congressionally mandated and independent studies show that the Title I program, even after 3 decades, has not diminished, much less eliminated, the poverty gap. Special education is comparable to Title I in federal spending, ineffectiveness, and inefficiency. It includes about a tenth of American children and currently costs $7.4 billion in federal money and an (imprecisely) estimated $35 billion to $60 billion, counting state and local contributions.

The Present Teaching Force

Maintaining certification as the criterion for employment and reemployment and graduate credits and experience as the basis of compensation may mean that unproductive teachers are paid just as much as their colleagues who best promote learning. These policies offer no incentives for improvement. Why should even the best teachers work hard and long when their compensation will be the same as the worst performers? Why not put their energies and talents into moonlighting, travel, or their families? A national survey of public school superintendents and principals corroborates such concerns. Large majorities of superintendents (76%) and principals (67%) said they need more autonomy to reward outstanding teachers. Almost the same percentages said they need more autonomy to remove ineffective teachers. Nearly all superintendents (96%) and principals (95%) said making it much easier to remove bad teachers—even those with tenure—would be somewhat or very effective.

Public-school teachers' salaries have long been chiefly determined by whether they are certified, their years of teaching experience, and their degree level. Despite thousands of doctoral dissertations in education written each year, little solid evidence shows these salary determinants promote student learning. In fact, studies by labor economists suggest that verbal ability, knowledge of the subject matter, and graduation from a selective college are at least as important as the usual salary determinants.

As for class-size reduction, in view of definitively inconsistent research and California's experience—where California policymakers spent about $5 billion per year from 1996 through 2001 to reduce class sizes in the first three grades, after which they could infer no achievement effect of class-size reduction—further class-size reductions seem unpromising. Such reductions, moreover, have been exceedingly costly.

Conclusions

Syntheses of experimental and quasi-experimental classroom studies of instructional methods and large-scale econometric studies reveal policies and practices that work well and cost relatively little. Other policies and practices, even though prevalent in American schools, are costly, but little evidence suggests their efficacy. Though more research would yield better estimates and resolve some uncertainties, the present body of knowledge about effects and costs suggests how American schools can be made more productive.

The Scientific Basis for the Theory of Successful Intelligence

Robert J. Sternberg, Yale University

Many different definitions of intelligence have been proposed. The conventional academic definitions are built around adaptation to the environment. But lay people's conceptions of intelligence seem much more linked to real-world success than are those of academicians. So, it may be useful to think in terms of the concept of successful intelligence, which deals not just with intelligence in its academic aspect but also as it pertains to all aspects of life. Successful intelligence denotes the ability to achieve success in life in terms of one's personal standards within one's sociocultural context. One's ability to achieve success depends on capitalizing on one's strengths and correcting or compensating for one's weaknesses.

The Theory of Successful Intelligence

Success is attained through a balance of analytical, creative, and practical processes, all aspects of intelligence, which enable individuals to adapt to, shape, and select their environments. These processes are applied to different kinds of tasks and situations, depending on what kind of thinking a problem requires. Analytical thinking is invoked when components are applied to fairly familiar kinds of problems abstracted from everyday life. Creative thinking is invoked when the components are applied to novel kinds of tasks or situations. Practical thinking is invoked when the components are applied to experience to adapt to, shape, and select environments.

A universal set of processes underlies all three of these aspects of intelligence. Metacomponents plan what to do, monitor things as they are being done, and evaluate things after they are done. Performance components execute the instructions of the metacomponents. Knowledge-acquisition components are used to learn how to solve problems or simply to acquire declarative knowledge.

Validation of the Theory of Successful Intelligence

The theory of successful intelligence has been internally validated by componential analyses, involving the information-processing components underlying performance on cognitive tasks, and factor analytic studies. The external validity of the theory of successful intelligence has been tested by correlational studies and instructional studies. The accumulation of evidence over some 25 years of research—including several studies with large samples that tested the theory with multiethnic, multinational subjects (e.g., 3252 students from the United States, Finland, and Spain in one study)—demonstrates that the theory of successful intelligence, encompassing analytic, creative, and practical abilities, provides a better prediction of success in life than does a theory comprising just the analytical element.

Improving School Achievement

Motivated by the belief that schools strongly favor children with strengths in memory and analytical abilities, we explored the question of whether conventional education in school systematically discriminates against children with creative and practical strengths.

We administered a test designed to measure specifically the analytical, practical, and creative abilities of 326 children from the United States and other countries who were identified by their schools as gifted. Children were selected for a summer program in introductory, college-level psychology if they fell into one of five ability groupings: high analytical, high creative, high practical, high balanced (high in all three abilities), or low balanced (low in all three abilities). Divided into four instructional groups, students used the same textbook and listened to the same lectures but differed in the type of instruction emphasized in their discussion section, which emphasized either memory, analytical, creative, or practical instruction.

Investigators observed that the students in the high creative and high practical groups were much more racially, ethnically, socioeconomically, and educationally diverse than were the students in the high-analytical group, suggesting that correlations of measured intelligence with status variables such as these may be reduced by using a broader conception of intelligence. Second, the investigators found that all three ability tests—analytical, creative, and practical—significantly predicted course performance. Third and most important, students who were placed in instructional conditions that better matched their pattern of abilities outperformed students who were mismatched. In other words, when students are taught in a way that fits how they think, they do better in school.

A follow-up study examined learning of social studies and science by third graders and eighth graders. The 225 third graders were students in a very low-income neighborhood in Raleigh, North Carolina. The 142 eighth graders were students who were largely middle to upper-middle class from Baltimore, Maryland, and Fresno, California. In this study, students were assigned to one of three instructional conditions: instruction with no intervention and an emphasis on memory; instruction emphasizing analytical thinking; and instruction emphasizing analytical, creative, and practical thinking. As expected, students in the successful intelligence (analytical, creative, practical) condition outperformed the other students in terms of the performance assessments. The result suggested that teaching for these kinds of thinking succeeded. Of greater importance, however, was the result that children in the successful intelligence condition outperformed the others

(continued)

even on the multiple-choice memory tests. In other words, to the extent that one's goal is just to maximize children's memory for information, teaching for successful intelligence is still superior. It enables children to capitalize on their strengths and to correct or to compensate for their weaknesses, and it allows children to encode material in a variety of interesting ways. In a third study, high-school students received reading instruction that sought to take advantage of their individually varying analytical, creative, and practical abilities. These students substantially outperformed students taught in standard ways.

Thus, the results of three sets of studies further suggest that the theory of successful intelligence is valid. Moreover, the results suggest that the theory can make a difference not only in laboratory tests but in school classrooms and even the everyday life of adults as well.

Improving Abilities

The kinds of analytical, creative, and practical abilities discussed in this essay are not fixed but modifiable: They can be taught. For example, one of our studies tested if it is possible to teach people better to abstract and generalize meanings of unknown words presented in context. Eighty-one participants were divided into five conditions, two of which were control conditions with no formal instruction. In the other three conditions, participants were taught either knowledge-acquisition component processes that could be used to abstract word meaning, the use of context cues, or the use of mediating variables. Participants in all three of the theory-based formal-instructional conditions outperformed participants in the two control conditions, whose performance did not differ.

Creative thinking skills also can be taught, and a program has been devised for teaching them. Investigators divided 86 gifted and nongifted fourth-grade children into experimental and control groups. All children took pretests on insightful thinking. Then some of the children received their regular school instruction, whereas others received instruction on insight skills. All children took a posttest on insight skills. The investigators found that children taught how to solve the insight problems using knowledge-acquisition components gained more from pretest to posttest than did students who were not so taught.

Practical intelligence skills can also be taught. One group of researchers has developed a program for teaching practical intellectual skills. Aimed at middle-school students, the program explicitly teaches "practical intelligence for school" in the contexts of doing homework, taking tests, reading, and writing. In a subsequent study of our own, this program was evaluated in a variety of settings. We found that students taught according to the program's approach outperformed students in control groups that did not receive the instruction.

Conclusions

Practical intelligence, like analytical intelligence, is an important antecedent of life success. Because measures of practical intelligence predict everyday behavior at about the same level as do measures of analytical intelligence (and sometimes even better), the sophisticated use of such tests could roughly double the explained variance in various kinds of criteria of success. Using measures of creative intelligence as well might increase prediction still more. Thus, tests based on the construct of successful intelligence might take us to higher levels of prediction.

The time has come to move beyond conventional and incomplete theories of intelligence and expand our notion of what it means to be intelligent. The "general factor" of intelligence postulated by older theories is overstated. Its generality depends on the populations of individuals tested, the types of materials with which they are tested, and the types of methods used in testing. Indeed, our studies show that even when one wants to predict school performance, the conventional tests are somewhat limited in their predictive validity. An expansion of the conventional conception of intelligence should include not just memory and analytical abilities but creative and practical abilities as well. A theory of successful intelligence fares well in construct validations, whether one tests in the laboratory, in the work place, or in schools.

Regrettably, children with creative and practical abilities, who are almost never taught or assessed in a way that matches their pattern of abilities, may be at a disadvantage in course after course, year after year. But the abilities underlying successful intelligence can be taught, and when classroom instruction addresses all of these varying and individual abilities, children have a better chance to learn.

The LSS REVIEW

STEPHEN PAGE, *Managing Editor*

DANIELLE SHAW, *Editor*

This publication is supported in part by the Mid-Atlantic Regional Educational Laboratory, the Laboratory for Student Success (LSS) at Temple University Center for Research in Human Development and Education (TUCRHDE) and through a grant funded by the Institute of Education Sciences of the U.S. Department of Education. The opinions expressed do not necessarily reflect the position of the supporting agencies, and no official endorsement should be inferred.

For more information, write LSS, Temple University, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122-6091, or call (215) 204-3000. Selected LSS publications and abstracts of articles, including back issues of *The LSS Review* and *The CEIC Review*, are available at http://www.temple.edu/lss/publibrary.htm

Science, Politics, and Education Reform: The National Academies' Role in Defining and Promoting High-Quality Scientific Education Research, 2000–2004

Lisa Towne, National Research Council/National Academies

This paper chronicles the National Academies' role in promoting high-quality education scientific research. It includes an overview of the current policy context, a description of two phases of work in education research quality undertaken by the National Academies in the past 4 years as well as related future initiatives, and a discussion of key issues that are likely to shape evidence-based education in the near term.

Policy Context

The central feature of the No Child Left Behind Act (NCLB) of 2001 is its testing and accountability provisions. Although compliance with these efforts continues to dominate state and local actions, the 111 references to "scientifically based research" (SBR) throughout the law have also started to garner the attention of policymakers. These SBR provisions in NCLB, in the Education Sciences Reform Act of 2002, and in parts of the pending reauthorizations of both the Higher Education Act and the Individuals with Disabilities Education Act set standards for the use of research to guide education policy and practice.

The SBR provisions are part of a broad-based, international push for "evidence-based practice," which can be traced to the 1950s and 1960s in medicine. The tools and applications of evidence-based practice extend beyond the United States, and several models integrating evidence to inform government decision making are in operation in many countries.

Major players in state and federal education posts do not agree whether the SBR provisions were intended to be, or are being interpreted as, strict legal requirements to adopt only programs with "scientific" backing or guidance to be followed as feasible. Either way, it is clear that education lawmakers and administration officials are promoting the widespread use of scientific evidence as a basis for decisions. Education research and its role in education policy and practice have been topics of discussion and the foci of reform efforts for some time, but the current prominence of these topics in federal education law and political rhetoric is unique. And the foregrounding of these topics has reignited age-old controversies about the nature of education research and its applicability to day-to-day education reform.

The Recent Role of the National Academies

The National Academies' operating arm, the National Research Council (NRC), has a 50-year track record of work in education research and reform. In 2000, the NRC's Center for Education initiated work focused on defining and promoting the quality of scientific education research. The first phase of this activity resulted in the publication in 2002 of the book *Scientific Research in Education*; the second extended this work by convening a series of high-level public forums for in-depth consideration of critical issues related to the quality of education research.

Phase 1: Committee on Scientific Principles in Education Research

In the summer of 2000, a bill introduced in a U.S. House of Representatives subcommittee to reauthorize the then Department of Education's Office of Educational Research and Improvement (OERI) included definitions of quantitative and qualitative methods for education research. The inclusion of these definitions signaled lawmakers' skepticism about the quality of education research and raised concern among education researchers that the law would have undue political influence over their profession. To infuse the perspective of researchers into such considerations, the NRC convened a multidisciplinary committee to articulate the nature of scientific research in education. The committee concluded that six common principles underlie all scientific inquiry, including scientific research in education:

- Pose significant questions that can be investigated empirically.
- Link research to relevant theory.
- Use methods that permit direct investigation of the question.
- Provide a coherent and explicit chain of reasoning.
- Replicate and generalize across studies.
- Disclose research to encourage professional scrutiny and critique.

The committee argued that although scientific research in education shared these principles with other disciplines, conducting such research was nevertheless different from doing so in other fields. Although not unique to it, certain features of education—such as the role of values and democratic ideals in schools; volition and diversity of people; and the variability of curriculum, instruction, and governance—are singular in their combination. The education research process thus requires close attention to powerful contextual factors. In addressing these and related issues, the committee also debunked many of the

(continued)

misperceptions about education research often implicit in policy debates. For example:

- Methods are the tools of science, not science itself.
- Theoretical frameworks play a crucial role in science, as does the need to identify, consider, and rule out plausible alternative explanations for observations. Research needs a skeptical community of investigators engaging in an ongoing, professional dialogue to consider how new theoretical and empirical findings fit into or challenge prevailing ideas.
- Methods themselves cannot be judged as "good," "bad," or even "scientific" absent the specifics of the inquiry itself.
- Randomized field trials are important tools for investigators pursuing causal questions in education, but no one methodology can adequately model or explain the complexities of education or any other area of inquiry.
- Hackneyed debates pitting quantitative methods against qualitative methods are not fruitful; both types of methods could be pursued rigorously.
- Research should be understood as a set of interrelated lines of inquiry conducted by multiple investigators over multiple years. Similar studies reaching different conclusions—common in a range of fields—are to be expected. The field of education would be more effective if it paid greater attention to the integration of studies, making them more than the sum of their parts.

Phase 2: Committee on Research in Education

The NRC anticipated the need for ongoing, structured dialogue among researchers, policymakers, and other stakeholders to enhance understanding of SBR in education and promote change that fosters high-quality education research. Therefore, it convened the Committee on Research in Education (CORE) to develop and implement events focused on important topics in promoting high-quality SBR.

CORE held a workshop series in 2003 and is issuing a set of related reports. The five events focused on pending policy and research issues and extended the central themes of *Scientific Research in Education*:

- understanding and promoting knowledge accumulation in education,
- journal practices in publishing education research,
- peer review in federal education research programs,
- implementing random assignment experiments in educational settings, and
- education doctoral programs for future leaders in education research.

The final product of the CORE project will be a report identifying the common issues raised during these events and outlining the committee's related recommendations.

Future NRC Initiatives

The National Academies will continue to promote improvements in education research as well as systematic connections between research and practice. Two initiatives are noteworthy in this context. First, the Strategic Education Research Project (SERP) will engage researchers and educators in a large-scale, long-term partnership to improve education research and its use. SERP is designed to develop the capacity and infrastructure for a sustained effort in linking education research and reform. SERP officials are currently working with university and state leaders to set up a system that will form the backbone of a fully functioning partnership. Second, the Division of Behavioral and Social Sciences and Education of the NRC is launching an initiative aimed at continuous improvement of social and behavioral research for policy and practice. This initiative will consider the connections between technical issues—such as evidentiary standards, theoretical and empirical lines of inquiry, internal and external validity, and replication and generalizability—and the use of research in the policy and practical worlds.

Looking Ahead

What will the evidence-based education movement, in which these NRC initiatives will be a part, look like in the coming years? The answer is, of course, speculative and multifaceted. Defining and enforcing standards of research quality will continue to be an important issue. While needed and appropriate, a broader consideration of the relationship between research quality—which is now viewed mostly as a technical matter—and its utility will be critical for the long-term success of evidence-based education. That is, concerns about evidence in education research must be framed in terms of its goal of utilization, with quality as a crucial but nonetheless supporting function.

One important vehicle for future discussions regarding education research quality and utility is the U.S. Department of Education's interpretation and enforcement of the SBR provisions across the many programs it administers. Consequently, the input of the research communities in government initiatives will be crucial to ensuring that quality standards are upheld in ways that capitalize on the full range of methods, perspectives, and strengths in the field. Additionally, self-policing of quality among education researchers through formal mechanisms—for example, manuscript review—and informal mechanisms will be less visible than those initiatives but more powerful.

Finally, efforts to focus attention on the implications of evidence-based education for educators will be needed. If evidence-based education is portrayed or understood as merely a federal mandate to implement off-the-shelf packaged programs that have been deemed "scientifically based," then it will likely be viewed as simply the most recent fad. Evidence can empower teachers and administrators to bring the best of what research has to offer to bear on their practice, to the betterment of all. Professional development across the continuum of education careers will need to embody this idea by integrating research and its application to the practice of education into its core.

American Board for Certification of Teacher Excellence: Applying Research to Develop a Standards-Based Teacher Certification Program

Kathleen Madigan, American Board for Certification of Teacher Excellence

The American Board for Certification of Teacher Excellence (American Board) represents a groundbreaking opportunity in education through its alternative teacher certification programs for prospective and veteran teachers. The American Board has two levels of certification—Passport to Teaching℠ certification, for aspiring teachers with a bachelor's degree in any field, and Master Teacher℠ certification, which recognizes experienced teachers for their exceptional subject proficiency and their students' strong achievement gains.

This paper highlights the American Board's Passport to Teaching℠ certification, emphasizing how this postbaccalaureate process for people interested in becoming teachers has been shaped by research: econometric studies showing that mastery of the subject matter is predictive of teaching success, experimental studies suggesting that some teaching methods are more effective in producing student achievement gains, and psychometric studies examining the development of high-stakes examinations.

The Importance of Teacher Excellence

Of all the strategies that hold promise for increasing student achievement, improving the quality of teachers will have the greatest impact, an impact that outweighs societal and demographic effects. Given the need for over 2 million new teachers in the next several years and the requirement under the No Child Left Behind Act (NCLB) for schools to have highly qualified teachers in core subject areas, many states are creating nontraditional or alternative routes to help interested individuals achieve teacher certification. Thus, it is important that while unnecessary barriers are removed to allow talented individuals into the classroom, the standards for becoming a new teacher must be increased. A standards-based approach to teacher certification could have a positive influence on teacher quality and quantity.

Researchers have established that teachers' academic competence and subject area proficiency correlate with student learning gains. Economists (as well as other researchers) suggest that the most reliable predictor for student academic achievement is how well the teacher performs on verbal ability tests. Economists have also found that high-school mathematics and science teachers who have a major in the subject area that they teach positively impact student achievement; these same studies suggest that having an undergraduate degree in the subject area has a greater impact on student performance than having traditional certification in those areas.

Research also shows that certain pedagogical strategies work better than others and that the pupils of teachers who use these methods are likelier to achieve higher academic performance. One large-scale research synthesis identified instructional variables that positively impact student achievement. Included in those variables were well-ordered classrooms and carefully structured instruction. Others studies have found that a positive disciplinary climate is directly linked to improved student learning.

History of the American Board

The American Board was founded in 2001with help from major reform-minded education leaders and a U.S. Department of Education grant. Less than 2 years later, the American Board launched a standards-based approach to teacher certification that includes preparation resources, online advisors, and a pioneering set of new computer-based teacher exams with nationally recognized passing scores. An additional 5-year grant was awarded to implement innovative teacher recruitment strategies, expand subject-area certifications, develop state-of-the-art preparation resources, create mentoring programs using technology, and evaluate the approach's effectiveness.

Cited in NCLB as one of the premier pathways to the teaching profession, the American Board has identified attributes of teachers and teaching that are currently measurable and that directly correlate with teachers' classroom effectiveness.

The Passport to Teaching℠ System for New Teachers

Passport to Teaching℠ certification is a career pathway for highly motivated, self-disciplined individuals interested in teaching who hold a bachelor's degree or higher from an approved college. The process is initiated by the completion of a pre-assessment of existing knowledge and skill mastery. Building from the base of that pre-assessment and working with their advisor and using rigorous standards, candidates create an individualized preparation plan that identifies key areas that require more experience or study. Once preparation is completed, candidates must pass two separate 4-hour, computer-based exams—the Professional Teaching Knowledge examination and a subject-area knowledge examination. Candidates have 1 year to complete the certification process. Once certification is achieved and the individual is employed, the American Board teacher is eligible for an online interactive mentoring program.

(continued)

Development of American Board Standards

The development of the American Board standards was comprehensive, representing a consensus of content specialists, including outstanding teachers, principals, administrators, scholars, teacher educators, researchers, psychologists, and policymakers recruited from diverse geographical regions, school sizes, and teaching experience representing diverse school populations. A seven-step process was used for standards development:

1. review of state and national teacher certification standards
2. analysis of highly regarded state K–12 standards
3. literature review and analysis of scientifically based research meeting rigorous requirements in linking effective teaching with student achievement
4. synthesis and distribution of documents summarizing findings
5. development of standards and framework
6. consistent review of drafts
7. revision and prioritization of standards during the development of the test blueprint

Overview of the Content Standards

A highly regarded group of panelists reviewed a large body of experimental and quasi-experimental studies to determine effective teaching practices connected to improving student achievement. The panel then synthesized the research to establish what a beginning teacher needs to know for classroom effectiveness. Standards were created and verified by educators and were organized around organizing, planning, and designing instruction for student success; effective instructional strategies; classroom management and organization; monitoring students and working with parents; and assessment.

The examination combines short-answer questions with interactive simulations that test candidates' readiness to enter the classroom and respond to classroom challenges. Because career changers are more likely to need training in professional teaching skills than in academic content, the American Board has developed extensive online resources to address this need. In addition to pedagogical concerns, tests for prospective elementary school teachers examine broad competence in the core elementary subjects. The middle-school and high-school mathematics examination tests for competence in all branches of mathematics applicable to the curriculum. The English examination tests candidates' knowledge of composition, critical reading, and literature. All candidates must also respond to an essay question designed to measure writing and communication skills; for the English teachers' essay, candidates must demonstrate their ability in literary interpretation.

Examination Development

Hundreds of educators participated in developing thousands of questions using several different formats. Each item was aligned with the test blueprint and the content standards—the hallmark of American Board examinations. Trial testing with more than 2,000 people was conducted. Panelists analyzed the data from the field-tested items and selected outstanding items to create the final examination. To maintain the security and integrity of the examinations, the American Board continually develops and field tests new questions.

Ideally, candidates are able to take the tests whenever they want to, alleviating the difficulties faced when trying to take less frequently administered state tests. Given the intended frequency and ready availability of the American Board tests, there will be several forms for each subject area. To further ensure the security of the test, candidates will only be able to take the test by going to an approved testing center. One especially powerful feature of the American Board examinations is that the test questions use multimedia formats to simulate teachers' daily dilemmas.

Establishing Nationally Recognized Passing Scores

In order to determine what constitutes a passing score for all four examinations, several different teams—more than 100 participants representing all aspects of education—gathered throughout the year to create performance-level descriptors and then analyzed the data using the modified Angoff rating system. To achieve Passport™ to Teaching certification, candidates must obtain scores at the proficient level (e.g., one standard deviation above the mean); to achieve Master Teacher certification, candidates must receive scores at the distinguished level.

Ongoing Research

The American Board will conduct a longitudinal study consisting of a descriptive analysis comparing the candidates and eventual awardees, an impact analysis of their success in producing measurable student gains, and an analysis of the duration of their employment in education.

Conclusions

The American Board is an undertaking of major and enduring national significance. It establishes the first nationally recognized alternative route to certification, making it both possible and practical for states to meet NCLB's mandate of having a highly qualified teacher in every classroom. By tapping into the large pool of skilled professionals who have the interest and ability to be highly effective teachers but who did not go through traditional teacher preparation institutions, the American Board is creating an efficient, comprehensive system to provide well-prepared and effective aspiring teachers with a pathway to U.S. schools and students, whose success depends so much on teacher quality.

The American Board also represents a critical step in reducing outdated regulations that create barriers, not standards, to becoming a teacher. By accepting American Board certification, states can focus on outputs, or verifying what teachers know and can do, and move away from overseeing inputs, or how knowledge and skills are acquired.

Evidence-Based Interventions and Practices in School Psychology: The Scientific Basis of the Profession

Thomas R. Kratochwill, University of Wisconsin—Madison

This paper provides an overview of evidence-based interventions (EBIs) and associated practices in school psychology. The profession has, for some time, embraced scientific principles and procedures across areas of professional practice, including diagnosis and classification, assessment, prevention and intervention, consultation, and research and program evaluation. More recently, the profession has embraced evidence-based prevention and intervention practices, intending to implement them in schools. However, doing so requires addressing multiple scientific and practice agendas, including preservice and in-service professional development, systemic school change to promote prevention and intervention program implementation, comprehensive models of mental health and educational services, and the sustainability of evidence-based practices.

Five issues need to be addressed for significant progress to occur in the evidence-based practice movement: (a) practice–research networks should be developed in school psychology; (b) intervention research methodology must be expanded to take into account practice contexts of EBI implementation; (c) practice guidelines could be developed to facilitate implementation of EBIs in practice settings; (d) professional development opportunities must be created for practitioners, graduate faculty, and researchers; and (e) collaborative partnerships must occur across the diverse groups involved in the EBI movement, especially those involved in generating the scientific database of EBIs.

Consideration of the scientific basis of school psychology interventions and practices is important because schools are the largest provider of child mental health services. Furthermore, growing evidence shows a reciprocal relationship between academic problems and disabilities and mental health problems. Thus, a scientific basis for school psychology prevention, intervention, and related practices seems essential to the promotion of students' academic success and mental health.

EBIs and Associated Practices in School Psychology

Following developments in evidence-based medicine, clinical psychologists developed a task force to review "empirically validated" treatments for child and adult mental health problems. The first clinical psychology task force report, released in 1995, stimulated considerable interest in other psychology specialty areas in addressing the EBI movement. The Task Force on Evidence-Based Interventions in School Psychology was formed in 1999.

The Task Force developed the *Procedural and Coding Manual for Review of Evidence-Based Interventions* for use in reviewing and documenting the research evidence for prevention and intervention programs. The manual's coding system is critical to establishing the scientific foundation of the field of school psychology, and reviews of intervention literature based on the manual's protocols are beginning to appear in school psychology journals. This application of these protocols and other work by the Task Force should also help to narrow the research–practice gap, that is, the disparity between what psychology and education research has revealed and the failure of that knowledge to impact significantly the work of practitioners. In addition, the Task Force has endorsed the idea that reviewing and documenting the evidence in support of interventions should be an ongoing, evolving activity and adopted the term *evidence based* to describe interventions judged to have credible scientific support.

Training and Practice in EBIs

The scientific foundation of school psychology can be evaluated by examining EBI practices in both graduate training programs and the practice of psychology in schools.

Graduate Programs

The Task Force surveyed graduate programs in school psychology to determine what they are teaching about EBIs, to investigate their integration of EBI training, and to understand any barriers to such training. Surveys were sent to 217 school psychology training directors, and 97 surveys were returned (44% return rate). The survey included a list of interventions already identified as evidence based by two divisions of the American Psychological Association (Society of Clinical Psychology and Society of Clinical Child and Adolescent Psychology), interventions determined to be effective by highly regarded scientific methods and empirically supported. Results of survey indicated:

- A relatively low percentage of school psychology graduate training directors were familiar with the EBIs included in the survey. When averaging across all interventions listed, 29% of directors reported being "not familiar," 30% reported being "somewhat familiar," and 41% reported being "familiar" with the EBIs.
- Exposure to the EBIs occurred more frequently in coursework than in practice experience. When averaging across all EBIs, 41% of directors reported that graduate students received "no exposure," 39% reported students received "exposure," and 30% reported students received "experience" with the EBIs listed.
- EBIs were rated as either "somewhat important" or "important."
- Lack of time was rated the most serious challenge to EBI training.

(continued)

- A high percentage of training directors reported that students were taught to apply the criteria developed by professional organizations in psychology and education when evaluating intervention outcome research.

A number of interventions considered evidence based by the training directors fell outside the EBIs incorporated in the survey. Some of these interventions have a weak evidence base.

No formal requirement within school psychology training programs mandates teaching EBIs, but the commitment to include scientifically supportable interventions in the curriculum will probably grow. Moreover, a number of graduate training programs embrace a scientist–practitioner model and are the most likely to embrace an evidence-based practice framework in future graduate training.

Evidence-Based Practices in Schools

Several recent surveys of evidence-based practices in schools do not paint a very positive picture. A study of the prevalence of substance abuse curricula in U.S. schools showed that many middle schools continue to implement curricula that are either untested or ineffective.

Another study investigated school psychologists' use of research in practice and the barriers to using research. Knowledge of effective intervention strategies and their use were closely matched, and respondents indicated they would like to use the strategies with greater frequency. Limited time was the top barrier to the use of all strategies. However, for cognitive behavior strategies and social skills training, practitioner training and the ability to adapt interventions to the school setting were significant factors limiting use; lack of support was indicated as a significant barrier to consulting with teachers, suggesting that some systemic support issues may be important.

Professional Standards and Influences on Practice

No formal requirements have made knowledge and use of EBIs and practice guidelines prerequisites of licensure and credentialing. The major professional groups involved in licensure and credentialing currently do not mandate this level of practice, and national school psychology organizations do not mandate EBI training as part of graduate program accreditation.

Many textbooks used in graduate school psychology programs and publications of the National Association of School Psychologists promote a scientific perspective. However, none of these, as yet, include the Task Force's evidence-based guidelines, though the guidelines have been distributed to training program directors and school psychology journal editors.

Promoting and Guiding the Use of EBIs in School Psychology Practice

Five strategies may promote EBIs:

1. Develop a practice–research network in school psychology.
2. Promote an expanded methodology for evidence-based practice that takes into account EBIs in practice contexts.
3. Establish guidelines that school psychology practitioners can use in implementing and evaluating EBIs in practice.
4. Create professional development opportunities for practitioners, researchers, and trainers.
5. Forge partnerships with other professional groups involved in the EBI movement.

The purpose of the strategies is to establish a link between research and practice that will help us better understand the effectiveness of interventions and promote their adoption and sustainability.

Looking Ahead: Barriers and Promising Trends

The study of graduate training programs revealed lack of time as one of the most serious obstacles to training in EBIs. More efficient methods of adding EBIs to existing coursework and enhancing faculty's skills must be found. When formulating competency-based training agendas, program organizers must thoroughly integrate field supervisors and other clinical faculty who are involved in direct supervision of school psychology graduate students. The 3-year curriculum of specialist-level training represents another time constraint. Many doctoral-level programs have more options for incorporating EBIs and related practices into courses.

A high percentage of trainers and students appear to be knowledgeable about the criteria developed by the Task Force for evaluating research. Increasingly, it will be important for graduate students to be exposed to the coding systems from various task forces and from the What Works Clearinghouse. Understanding these criteria will promote understanding and selection of appropriate EBIs.

It will also be important to examine not only interventions and prevention programs identified as evidenced based by the task forces but also other interventions and programs with a strong educational and prevention focus. Professional groups must disseminate information to school psychology trainers to help them select EBIs. Students who receive EBI instruction in graduate school should master these programs within a competency-based framework, ensuring that students acquire the skills in a practice context.

Finally, a promising direction in establishing EBIs in school settings is adoption of multiple levels of intervention programs. Three-tiered systems of prevention are promising because students can progress through a series of interventions before receiving traditional services such as special education. It is critical to teach faculty and graduate students strategies for systemic change in schools so that such systems can be adopted. Such content will facilitate the adoption and sustainability of evidence-based practices and interventions.

The Institute of Education Sciences' What Works Clearinghouse

Robert Boruch and Rebecca Herman, University of Pennsylvania

The What Works Clearinghouse (WWC), established in 2002 by the Institute of Education Sciences (IES) of the U.S. Department of Education, was designed to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence on what works in education. The focus of the WWC is on the evidence pertaining to the effects of interventions, notably evidence that permits causal inferences. The WWC does not endorse particular interventions nor conduct randomized trials or quasi-experiments; rather, part of its mission is to assure that all reports on such studies in a selected topic area are identified and screened for dependability of the evidence.

This article outlines the main features of the WWC as of May 2004. Because the effort is evolving, and changes are made when we see opportunity for improvement, readers are encouraged to consult the WWC's website, http://w-w-c.org, for up-to-date information.

Operating Principles and Organization

Assuring the quality of evidence is an operating principle, represented partly in the WWC's focus on scientific standards. The WWC embodies scientific standards in at least three ways: It seeks unbiased estimation of an intervention's effect; it applies methodological advances in meta-analysis, reviews, and standards that have been developed for assessing assemblies of studies and reporting systematic reviews; and its processes are overseen by an independent Technical Advisory Group (TAG) and a peer review system.

Because of its focus on scientific excellence, the WWC particularly values randomized trials and, albeit to a lesser degree, good quasi-experiments. A second operating principle requires the WWC to be procedurally and organizationally efficient. Because the WWC is exploring new terrain, a willingness and capacity to improve is a third operating principle. Emphasizing accessibility and transparency in organization and procedures, in identifying and explaining the evidential standards, and in efforts to improve constitutes a fourth operating principle under the contract.

The WWC's Topics and Reviews

The WWC first selects a particular topic, based on suggestions from any individual or organization; the selection depends on (a) the relevance of the topic to current education policy and practice, (b) the topic's probable importance in decisions about what interventions can be adopted, and (c) the level of evidence available. The initial topics for review include middle school math (now available on the WWC website), peer-assisted learning, dropout prevention, adult literacy, character education, beginning reading, reduction of school violence, and English language acquisition.

A WWC review on an intervention topic begins with a detailed, publicly accessible protocol that defines the intervention and inclusionary criteria, the target population, the outcome variables that are pertinent, and the study designs that are eligible or ineligible for a WWC review of any kind. An extensive literature search identifies studies of interventions within the defined topic area, and randomized trials and high-end quasi-experiments are admitted to candidacy for WWC review.

The review process then depends on double coding—coding by two independent coders—of certain characteristics of each category of study that influence internal validity. For instance, a randomized trial that has a large difference in the attrition rate between intervention arms could be downgraded to quasi-experimental status absent other information that speaks to the biases that attrition engenders. A series of codes that, in effect, say, "Yes, attrition occurred and is potentially a problem" is tied to the parts of the study narrative that cover attrition.

The WWC Study Report

All this leads to a WWC Study Report on a particular piece of research on an intervention's effect on a particular target, in a particular context, and reported perhaps in different ways or series of articles in peer-reviewed journals or by a private organization, commercial publisher, or other vendor. The WWC's Study Report contains a synopsis of the study of the effects of an intervention and a summary of its strengths and weaknesses relative to WWC's uniform standards of evidence, which are described below. A WWC Study Report, for example, might report on a study of a randomized trial of a particular school-based truancy program. WWC Study Reports undergo a series of reviews, including one by the study's authors, before being made public.

The WWC Intervention Report

The second level of review results in a WWC Intervention Report and depends on all relevant WWC Study Reports concerned with certain class of interventions, for example, school-based truancy programs. The WWC Intervention Report describes the intervention, summarizes the studies that were reviewed to understand the intervention's effects, and explains why their evidence is dependable. For instance, an Intervention Report might synthesize results of a half dozen Study Reports, each of which covers a specific form of school-based truancy program. The

(continued)

Intervention Reports' conclusions, including key findings, their generalizability, and gaps in the evidence are reviewed independently by methodologists, the WWC Steering Committee, and the independent WWC Technical Advisory Group.

The WWC Topic Area Report

WWC Topic Area Reports, a third level of review, will cover different interventions that are supposed to affect similar outcome variables in specified target populations. All interventions on a given topic, such as truancy, are considered, and all studies done on each class of interventions are reviewed. A Topic Area Report on truancy interventions, for instance, might then cover all interventions classified as "school-based interventions," "court-based," and "parent-oriented," and others. The WWC Topic Area Report covers all WWC Intervention Reports that fall in the ambit of the topic. The resulting evidence is evaluated and summarized.

Standards of Evidence

A major theme underlying all standards enunciated by the WWC is that one must be able to make causal inferences about what works, what does not work, and what harms. As a practical matter, this means that all report standards pay attention to randomized trials, to certain types of quasi-experiments trials, and to the important differences in dependability of randomized trials as opposed to the important forms of quasi-experiments.

The standards for inclusion in a WWC Intervention Report are embodied in two protocols that have been vetted repeatedly and publicly in a variety of forms; the results of vetting are given on the WWC's website. The first, the Design and Implementation Assessment Device (DIAD), attends to about 40 characteristics of individual studies that are targeted for review in WWC Study Reports. The DIAD addresses four kinds of validity in designing and executing studies of an intervention's effects. The first concerns construct validity, that is, the extent to which an intervention is well defined, relevant outcome variables are well described and measured, and outcome measures are reliable. Studies that pass muster on these accounts are tentatively admissible for an Intervention Report.

The second standard in the DIAD evaluates internal validity and focuses on studies designed to produce statistically unbiased estimates of relative effects, that is, randomized trials in particular. Quasi-experimental designs, such as regression-discontinuity, are tentatively admissible. A series of standards also evaluates the design's execution. A well-designed trial that is executed in a way that does not compromise the design would be highly rated. Threats to internal validity, such as appreciable differences in attrition among the arms of the trial, have to be recognized and dealt with by the study.

The third broad DIAD standard is that the target samples, including important subgroups, settings, and outcomes, must be pertinent to the topic and intervention under review. If the studies' sample is so idiosyncratic as to permit no generalization to an important target population or the sample is not specified well, the study's results are rated at a lower level of quality.

The final broad standard concerns statistical analysis. Operationally, this standard requires that a study meet conventional statistical assumptions, especially about independence of the statistical variations (error) in the sample. Statistical data that permit estimation of effect size, and relevant sample sizes are required for studies to receive high marks.

The second major standard for evaluation is the Cumulative Research Evidence Assessment Device (CREAD), a tool for summarizing the totality of evidence from multiple studies and assessing its credibility. The CREAD depends on scientific work over roughly the past 2 decades in health care, criminology, and welfare, as well as education that seeks to understand how to summarize the results of studies uniformly and against clear standards.

WWC Evaluator Register

The IES's WWC plans an Evaluator Registry that provides information about people and organizations that have the capacity to produce high-quality evidence on the effects of educational interventions. Information about who and what organizations produce high-quality research is potentially important to school districts and publishing firms, for instance, that do not themselves have the capacity to generate evidence that meets high standards.

Conclusion

The WWC aims to assure that its products are confidently used by policy people, practitioners, researchers, and others. The WWC seeks to accomplish this goal through its unprecedented focus on the quality of evidence that is generated about the effects of education interventions and its focus on scientific standards in making judgments about evidence quality, all through a process that is public and as transparent as possible.

Conclusions and Recommendations

Herbert J. Walberg, University of Illinois at Chicago, and Rena F. Subotnik, American Psychological Association

The conferees met seven times to set forth conclusions and recommendations based on the preliminary, pre-circulated versions of the chapters and the participants' own expertise and experience. Smaller work groups met several times to formulate consensual recommendations to present at the last plenary session of the conference. Not every participant agreed on every point, but several points gained consensus. This concluding chapter summarizes largely agreed-upon recommendations as well as dissenting views. We have felt free to consolidate duplicative recommendations from the separate small group notes, reorganize the material, use our own words, and explain points that might seem overly terse outside the conference deliberations.

Quality of Research

The standards of education research should be raised, and research should better address important questions of education policy and practice. Randomized control-group experiments are most definitive in probing causal assertions, which should serve as one of the bases of K–12 education practice and policy along with costs and other practical considerations. Other forms of research, however, can supplement and complement randomized experiments. Compared with randomized experiments, they have some distinctive advantages and disadvantages worth enumerating and considering:

1. *Quasi-experiments*: These designs employ preconstituted groups and are usually much cheaper and easier to do than experiments because they do not require arbitrarily reassigning students to groups. For this reason, they may also enable investigators to make their findings more generally applicable by studying different kinds of students in a great variety of school conditions. Quasi-experiments may also be more realistic because, by their nature, experiments are contrived and may lead to Hawthorne or "hothouse" effects. But quasi-experiments may be less causally definitive than randomized experiments because the groups may differ substantially in various known and unknown ways before the study starts.

2. *Formative research*: Engineering and garage experimentation, as in the case of the first Apple computer, follow a rich tradition of pragmatism that remains vibrant today. The Wright brothers didn't employ randomized experiments leading up to the first human flight. As they tinkered with various wing configurations and other plane variations, they gradually added improvements as they gained information from failures as well as successes.

Similarly, behaviorists have long carried out rigorous research on one individual at a time and shown sharp differences in behavior between alternating experimental control periods. With this design, large and obvious effects even make statistical inferences unnecessary. Particularly in the development of computer-based instruction programs, such "formative research" rather than experiments is in order. But these programs may require later experimentation, independent of the developers, to prove their efficacy.

3. *Observations*: Observations can lead to fruitful hypotheses for testing in experiments and other research designs. Reflecting a long-lived precedent in science, "outlier studies" of exceptionally high- or low-performing individuals, organizations, or even countries may prove particularly fruitful. For example, the famous 1983 report, *A Nation at Risk*, stimulated interest in Japanese schools, which produce high achievement at comparatively low costs. The report's background papers and subsequent research on Japanese schools revealed features that explained their productivity: intensive maternal support of their children's studies, a school year of 240 days in contrast to the usual 180 in America, a nationwide curriculum, competitive examinations for admission to middle and high schools and college, the prevalence of private tutoring schools, and knowledgeable teachers.

Observations of experiments themselves may also be fruitful particularly to investigate whether or not programs have been well or poorly implemented. Perhaps, for example, a new practice or policy was poorly applied or even remained largely unapplied, which might account for no differences among experimental and control groups. Even well-demonstrated practices may require observation to monitor how well they are being used. But observational research usually cannot stand on its own in making causal inferences. Aside from possible observer bias, observations are costly and therefore limited in number and generalizability.

4. *Regression analysis*: Economists and policy analysts have a long tradition of inferring causality from analyses of non-experimental data, which they employ partly because they cannot, for example, randomly change currency values and tax policies. They can, however, include and test rival hypotheses in regression equations to test their validity.

Still, their analyses may depend heavily on theories and assumptions that lack evidence and consensus. For this reason, economists and policy analysts increasingly turn to experimentation when policy questions are sufficiently important such as in welfare reform and job

(continued)

training. They can also make use of "natural experiments," such as in the case of oversubscribing students lotteried in and out of charter schools and voucher programs.

5. *Cost effectiveness*: Even statistically "significant" programs may have such small effects, high costs, or implementation difficulties that they make a poor choice for continuation and expansion. A "decision tree" may be helpful in making explicit such considerations when coming to major decisions. Thus, the size of policy, program, and practice effects should be weighed against but not necessarily dominated by costs, including training and other requirements. Chosen programs should be both efficient and effective.

6. *Consumer research*: Several conferees held that educators' and clients' opinions should not be designated as rigorous or even relevant since they may not be well informed about risks and outcomes. Still, many human contacts in free societies involve voluntary transactions determined by lay opinions and partial knowledge. Thus, in evaluating charter schools, investigators, policymakers, parents, and others might greatly value information about both achievement effects and parental satisfaction.

7. *Synthesis research*: In education, a single study should rarely be the sole basis for conclusions, recommendations, and decisions. For this reason, the What Works Clearinghouse holds great promise for meta-analyzing many high-quality studies to come to relatively definitive findings, particularly about the size and consistency of achievement effects. Summaries and critical reviews of studies are also valuable, particularly if they can reasonably conclude that findings from a variety of studies in several of the categories above lead to the same conclusion. Actually, a substantial corpus of such extant research is ready for such syntheses both for application to K–12 and as a basis for future research.

A single investigation may combine two or more of these methods. Observations, for example, may illuminate experiments; cost data may be simultaneously gathered to inform decision making. In addition, programs of sequential research may efficiently yield great benefits. Syntheses, for instance, may suggest hypotheses that lead to formative studies of an idea, followed by an experiment to test its field efficacy in ideal circumstances, followed by consumer research and quasi-experiments to probe its attractiveness to users and its effectiveness for various students in a variety of circumstances.

Research Questions

Before collecting data, some questions are generally applicable to many areas of education research. They are:

- What is the problem? What is its nature, severity, and context?
- Who thinks so? Why? Have they thought comprehensively?
- What considerations may have been omitted or slighted?
- How does the problem affect service delivery and outcomes?
- What are the exceptions to the problem? Do they suggest solutions? What do professional judgments suggest?
- What does extant research indicate? How well? What are the gaps in knowledge?
- What are the most promising solutions to be investigated?

These questions can be even more important than the choice of methodology and rigor of investigations: Better an approximate answer to right question than a precise answer to the wrong question. Clear questions, moreover, can lead to clear conclusions. And the means of investigation should be guided by the questions asked and the body of extant knowledge available.

Question Origins

Questions might originate with policymakers, educators, or investigators. For example, the field of educational psychology may be thought of as chiefly deriving principles from psychology and applying them to educational practice as motivated by problems and questions posed by policymakers and educators. It is a tricky and difficult business: We all seem cursed by insufficient time, and psychologists and educators (and undoubtedly others) do not ordinarily read each other's literatures and may not even speak to one another. The field of psychology is splintered; hence, educational psychologists themselves pursue such narrow specializations as measurement, motivation, and instruction. Like other academics, educational psychologists may tend to know more and more about less and less.

A personal anecdote to illustrate the problem even within a single organization: A former U.S. secretary of education asked one of us (Walberg) to investigate the possible relation of policy and research in the U.S. Department of Education. Interviews with the assistant secretaries and the staff of the National Institute of Education, then the research division, and of the division heads of elementary-secondary, higher, special, and bilingual education revealed little correspondence between the research questions being pursued and the questions policymakers were asking. What to do?

Possible Research-Policy-Practice Links

Actually, professional organizations that include more than 2.5 million members do or potentially can reduce the research gap. They include the two very large teachers unions as well as such member groups as the National Association of Secondary Schools, the Education Commission of the States, the Association for Supervision and Curriculum Development, and the Educational Leaders Council. In addition to local, state, and national conferences, they publish books, magazines, and pamphlets and disseminate research and ideas.

The U.S. and state departments of education and larger school districts make research information available. The American Legislative Exchange Council, the Brookings Institution, the Hoover Institution, and the Heartland Institute carry out and make available policy research with Congress and state legislators as their intended audience.

Yet the gap remains, and in some instances, educators and their organizations have promoted policies and practices based on inadequate and non-independent research. What can be done? What new and provocative ideas should be considered?

Education policy analysts sometimes look to the medical model. Though it took centuries to accomplish, physicians are educated not only in evidence-based procedures but in their evidentiary basis. Increasingly, they can draw upon broad meta-analyses of many studies conducted throughout the world. For problems for which they are unfamiliar, they can make use of such publications as the *Merck Manual* and more specialized handbooks on diagnosis and treatment.

In medicine, multiple randomized experiments with placebos and other features help enable confident causal inferences. When experiments cannot be performed, physicians and public health officials can base practice and policy on the work of biostatisticians and epidemiologists who carry out statistically controlled studies as in econometric research. Physicians and hospitals can be sued for malpractice or violating the evidence-based standards of practice. Increasingly, hospitals and physicians are rated by various consumer organizations. Medicine is increasingly incentivized and sanctioned.

Still, medicine is different from education in many ways, including its longevity as a scientific field, the size of research funding, and the constraints faced. In addition, pharmaceutical firms sponsor medical research, which may not be entirely independent and objective, especially when researchers work for or own firms. Holding medicine as the model for education argues by analogy rather than evidence of success. Skepticism and open-mindedness seem in order.

Even so, it seems reasonable for undergraduate majors in education, like physicians-to-be, to know not only about research conclusions but also how evidence is gathered and analyzed; they should become critical consumers of research. Graduate programs might foster completion of small studies and the critical syntheses of research on a given topic. One working group of conferees went so far as to say that teacher tenure and merit raises should be granted in part on syntheses and successful applications of research.

Another provocative and possibly useful analogy is the behavior of business firms and markets. The problem in education is "disseminating" research. The problem for firms is retaining proprietary trade secrets. Why? Firms are driven by market competition; they conduct formative research and market surveys; they try to favorably "brand" their products and services.

The federal No Child Left Behind Act and state legislation lead to closing failing schools and opening up competition to charters and other forms of privatization. Under such régimes, for-profit educational management organizations such as Edison Schools carry out substantial formative and experimental research on the effectiveness of their methods. They are more likely than conventional public schools to carry out market research and "brand" their offerings. Such research may create tighter links between research and practices than employed in public schools. Again, open-mindedness and skepticism are in order.

Know That, Know How, Can Do

To increase knowledge utilization, investigators and educators must somehow collaborate or at least communicate. But how? A 1960s answer was dubbed "action research," a strategy in which educators themselves simultaneously did research and put it into practice. Perhaps because of the difficulty of doing two things well and the increasing division of labor in modern societies, action research fizzled and was characterized as being neither action nor research.

In view of the continuing difficulties of research collaboration, the Laboratory for Student Success, one of the conference sponsors, adopted at its inception the motto, "Know That, Know How, Can Do." This motto is intended to suggest the acquisition and knowledge of evidenced-based principles, the general knowledge of putting them in practice, and, distinctively, how practical educators can suit them to their own purposes, students, and conditions.

The conference itself exemplifies the motto's application. Eminent scholars from around the country write authoritative "Know That" chapters summarizing the principles from an area of research, in the present case methods of research themselves. At the conference site, they meet with parents, educators, leaders of Washington-area organizations, and federal agencies to discuss "How To"—or how the principles can be generally employed—as well as "Can Do"—or how the principles can best be suited to the educators' particular students, purposes, needs, and circumstances.

All three sets of conference ideas are shared, first, in a promptly published journal, *The LSS Review,* which is simultaneously made available on the LSS website and which may be freely accessed and downloaded by anyone interested in the topic (http://www.temple.edu/LSS). Later, taking into consideration the discussion of policy leaders and educators' concerns and insights expressed at the conference, the authors revise their educator-informed "Know That" chapters for the published book. Subsequently, further conferences take place, led by one or more of the original conferees and geared even more to educators and policymakers informed by "Know That" but focused more on "How To" and "Can Do."

The Laboratory for Student Success at
Temple University Center for Research in
Human Development and Education
1301 Cecil B. Moore Avenue
Philadelphia, PA 19122-6091

Non-Profit
U.S. Postage Paid
Philadelphia, PA
Permit No. 1044