Denis Newman
Michael Cole

Can Scientific Research From the Laboratory be of Any Use to Teachers?

Behavior in a psychology laboratory—constrained by the need to efficiently replicate tasks, record individual responses, and avoid contamination from external factors—is different in systematic ways from behavior within an everyday environment where similar tasks are undertaken and problems solved. This article describes a program of research that identified the sources of this "ecological invalidity" of laboratory settings. The authors connect these insights to current attempts to apply laboratory controls in field research in schools where measures of the effectiveness of instructional programs can be based on high-stakes testing. While recognizing important applications of controlled experimentation both in the laboratory and in the educational policy research, they also find potential for the experimental controls themselves to lead researchers and decision makers to the wrong conclusion.

For some decades, we have been trying to understand how the teaching and learning process works and why some populations of children don't seem to learn as well or in the same way as others. Trained as experimental psychologists, we relied heavily on experimental methods to test theories of the processes said to underpin successful learning and teaching. However, our focus on population differences in cognitive performance brought us into territory unfamiliar to many psychologists, namely the practice of teaching and how children think when they are not being taught or tested. We found a gap between theories of learning and development that emerged from experiments on individual children and the classroom teaching and learning practices to which those theories were supposed to speak. Psychological research conducted under well-controlled conditions did not clearly map onto the complex ecology of schools and the other settings where children grow up (Newman, Griffin, & Cole, 1989).

This gap between laboratory-style research and actual practice continues to stand in the way of current attempts to apply scientifically based research to improvements in practice. The very constraints that make it possible to carefully control conditions in an experiment and attribute performance to individual children can leave the researcher blind to how behavior in such circumstances is differently constrained and enabled by the ecology of actual classroom practices. Our goal in this article is to explain why this gap exists and to describe how we attempted to make the gap itself the subject of our scientific research. We can

Denis Newman is president of Empirical Education Inc., Palo Alto, CA; Michael Cole is a professor of communications at the University of California, San Diego.

THEORY INTO PRACTICE, Volume 43, Number 4, Autumn 2004

then apply our findings to current proposals for using laboratory methods to improve educational practice.

Laboratories and the World Outside

Our first venture into exploring the gap between experiment and classroom occurred in central Liberia in the 1960s. Michael Cole was sent to West Africa to help figure out why local children seemed to have such a difficult time learning mathematics, even very elementary mathematics, in school. Owing largely to a lack of training, he approached this problem by assuming that in order to understand children's mathematically relevant understandings in school, it would be useful to know about their mathematically relevant experiences outside of school. Thus began a career in comparative human cognition (Cole, 1996) and to the conclusion that both standard psychological testing and schooling are systematically different from the everyday practice of the behaviors that they purport to describe, measure, and assess in many ways. In psychology this is known as the problem of "ecological validity."

The problem we are addressing is not new. Discussions of ecological validity in psychology date back to at least 1943 and the debate between Brunswik (1943) and Lewin (1943) on ecological psychology versus psychological ecology. Powerful contributions to the idea of the limitations of laboratory methods for psychological analysis have been contributed by Bartlett's (1958) distinction between closed systems and everyday tasks, Gibson's (1966) work on perceptual systems, Neisser's (1976) discussion of cognition and reality, and Bronfenbrenner's (1979) notion of experimental ecology (Cole, Hood, & McDermott, 1979). These discussions were not calling for the elimination of laboratory experimentation. However, they resulted in important cautions about accounting for the systematic limits that experimental controls put on the generalizability of the findings beyond the conditions under which they were obtained.

Comparing settings within classrooms

During the 1970s, with our colleague, Peg Griffin, we began working on the problem of how to describe this gap between methods of experimental research and what goes on in everyday settings (Newman, Griffin, & Cole, 1989). Our main question was: Is it possible to claim that children who do poorly in testing situations can, nevertheless, do well in logically equivalent everyday settings? Can a task presented in a formal test predict achievement on the same task when encountered in everyday life? At the outset, it was far from clear that we could even identify a cognitive task outside of the confines of the laboratory. To make this problem tractable, we decided to compare how tasks emerge in formal and less formal events within the confines of a classroom.

Over a period of 2 years, we collaborated with teachers in a third- and fourth-grade classroom to create full curriculum units in subject areas such as social studies (Native Americans before extensive European contact), chemistry (chemical properties and reactions), and math (long division). Each unit was a research cycle where we identified a specific task embedded in a one-to-one tutorial (to represent the traditional laboratory test), in small group activities, and in whole group lessons. We collected video records of all these classroom events because we saw each as an interactive construction to be inspected in detail for how or whether the task appeared.

We are using the term *task* to refer to the kinds of things that psychologists ask children to do during laboratory experiments. This might be something such as, remember a list of nonsense syllables; solve a math problem presented on a computer display; or solve a puzzle presented in the form of familiar objects such as half-full glasses of water, playing cards, or skits recorded on video. Usually there is some goal to the activity and the point is to get the child to respond in a way that can be quantified or at least categorized by a research assistant.

For example, in our chemicals unit we embedded Piaget and Inhelder's (1975; Inhelder & Piaget, 1958) famous combinations task where children were asked to find all the pair-wise combinations of a set of objects. In a one-to-one tutorial, we presented the children with the task of finding all the ways that a set of six movie stars could go together. The classic solution to this problem is what Piaget called "intersection"—the child creates a

conceptual matrix to generate all possible pairs. A less sophisticated method is to make up pairs until "you can't think of any more," which lacks the certainty of intersection. The tutor (our research assistant) carefully calibrated her scaffolding of the task so we could tell whether the children could solve the problem on their own and, if not, to introduce hints and, finally, more explicit help until the child implemented the intersection procedure.

In this case, the task was to find all possible pairs. The outcome measure, in the Piagetian framework, was whether the child used the intersection procedure unassisted (indicating a certain abstraction of thought) or whether the child used a strategy characteristic of a lower stage of development.

This is the task we wanted to confront the children with outside the rigid control of the tutorial session with our research assistant. We wanted to see if we could set up a situation where the children would take on the task of finding all the pairs without having an experimenter actually explain the task explicitly. Later, as part of the chemistry unit, groups of 3-4 children worked on their own with beakers of sodium meta-bisulfate, Clorox bleach, copper sulfate, and potassium iodide—each pair producing a distinctive reaction. The teacher instructed them to fill out a worksheet to record what happened when they combined all the pairs of chemicals. Students displayed their own varied motivations to find as many ways as they could to mix the chemicals. None of the groups, however, started the activity by setting out to find all the pairs. In other words, the task they had been asked to solve in the laboratory was not immediately apparent to them. It was only when it came to checking to see if there were any more pairs to mix, that the task we had sought emerged in some groups. There was only a mild correlation between the appearance of the intersection procedure in the children's interactions and the individuals who were most successful in using the intersection procedure in the tutorial.

What these observations say about laboratories and development

Evidence from this unit and others showed that in some well-designed settings it was possible to identify the performance of a cognitive task outside of a laboratory. But the differences were substantial. Outside of the laboratory, the children must find the task as a solution to an emergent problem. If the task does emerge, it is often difficult to identify which member of a group was responsible. As a result we found it very labor-intensive to score the data even with full video records shot from two angles. In contrast, laboratory tasks are designed for precise replication and easy recording and scoring of the response. The task is *presented to the subject,* never discovered by the subject. Unlike informal or everyday settings, the laboratory is designed to replicate with fidelity a task or treatment.

As indicated previously, our work was planned as an investigation of the gap between theory and practice rather than a test of a developmental theory. We were, however, greatly assisted by, and contributed to, the growing scientific literature showing how development of cognition proceeds from concrete, external activities involving the support of other people, to abstract, internal processes (Cole, 1996; Rogoff, 2003; Vygotsky, 1978; Wertsch, 1986). Originating with Vygotsky's concept of a zone of proximal development, a fundamental insight offered by this theory is that children can perform tasks with help from others (i.e., scaffolding) before they can do the same thing on their own. This is a view that is now part of the mainstream scientific consensus in learning theory (Bransford, Brown, & Cocking, 1999)

Classrooms and everyday settings can potentially provide a wide range of supports for thinking and learning. Viewing development as primarily a social process helps us understand the classic laboratory task as an odd kind of interaction where the subject is helped to find the task (by being presented with it) but provided no help in actually solving it. While uncomplicated to replicate, from the point of view of the classroom, such tests are time taken away from the core processes of teaching and learning where a teacher's continuous assessment is far more problematic.

The Current Context for Moving Developmental Theory to Practice

We are writing this article in interesting times for educational theory and practice. The No Child Left Behind (NCLB) Act of 2001 requires that

school decision makers take account of "scientifically based research" in their choice of instructional programs purchased with federal money. In tandem, the U.S. Department of Education's research unit has been reinvented as the Institute for Education Sciences. Its role model is the National Institutes for Health (NIH), which sponsors scientific research aimed at solving problems in medicine and related fields. Grover Whitehurst, director of the research unit, explains: "My marching orders are to fund research that is scientifically strong, that is relevant to pressing problems in education, and that will be utilized by educators and education decision makers" (2002).

We applaud the interest in objective evidence as the basis for improvements in educational practice. But like many researchers working in education, we are also cautious of attempts to link laboratory findings and methods directly to policy and procurement decisions as is now mandated by federal law. The American Educational Research Association, in its journal *Educational Researcher,* has provided a forum for this debate and we will not attempt to summarize all the issues (Slavin, 2002; Jacob & White, 2002). However, our earlier work examining the gap between theory and practice leads us to address one issue in particular: the difficulty of replicating experimental treatments on a large scale.

Medicine as a model for education

In modeling the new Institute for Education Sciences on the NIH, medical research is held up as an icon of excellence for educators. Speaking at a seminar hosted by the U.S. Department of Education, Stephen Raudenbush (2002) makes the case for the strong parallel between medicine and education, pointing in particular to the similarity of the debate in medicine more than 40 years ago and the debate today in education. The debate 40 years ago concerned the need for large-scale clinical trials to establish that a new vaccine, a new surgical procedure, or a new medication was more effective than current practice. Some felt that the "cold logic of science should not replace the clinical judgment of the seasoned practitioner." Ethical objections were raised about withholding drugs or procedures from randomly selected control groups when it could save a life. But now the consensus is that large-scale clinical trials, while not foolproof, have been beneficial for medical practice.

In the U.S. Department of Education's definition of scientifically based research (commonly abbreviated SBR), much is made of the procedure of random assignment of subjects to the experimental treatments and the control groups (Mosteller & Boruch, 2002). Random assignment is standard procedure in scientific experiments because it is the best way to assure that the deck is not stacked for or against the experimental treatment (Shadish, Cook, & Campbell, 2002). It assures that no characteristic of the subject that might affect the efficacy of the new treatment (even one we don't know about) is represented more in one group than the other. In doing so, the variability in response to the treatments can be assumed to be evenly distributed, allowing researchers to use statistical methods to decide if the positive effect of the treatment can rise above the noise of the randomly distributed variability. If it does, then researchers can be confident that it was the treatment that caused the positive effect. Students in any experimental science learn this principle in Introductory Statistics.

The focus on random assignment of subjects in clinical trials as the model for educational research unfortunately makes both medicine and education sound easier than they really are. By focusing on the human trials at the very end of the process of approving a medical procedure, the hard work of understanding the underlying biological processes and the creative work of inventing the new treatment can be overlooked. Medicine is not as simple and clean as the cold logic of statistical methods may appear. Also, while the statistical analysis of the experiment can give researchers confidence that a new medicine caused an increase in recoveries, it says nothing about *how* the medicine caused the effect. Often the underlying biological processes remain a mystery embedded in complex interactions about which medical science has yet to develop a consensus. The next level of improvement of the treatment awaits this further research.

Parallel issues arise in experimental research in educational settings. As Erickson and Gutierrez (2002) point out, given experimental evidence that an educational innovation caused an improvement,

we still don't know *what* actually caused the improvement. Qualitative research may be needed to understand how the effect was achieved. What is in the new program that accounts for the beneficial changes in the classroom? Was it the new content of the program or sequence of presentation? Was it the careful scripting of the teacher's interactions or simply greater time-on-task? Was it the textbook's new approach to the topic or simply the excitement of having brand new glossy pages? Without knowing the immediate cause of the improvement, we don't have a theory that helps us design the next, even better, program.

This is not to say that experimentation should be eliminated from educational research. In fact, as Cook (2002) points out, there are many instances where well-designed experiments should have been used but were not. In many cases, the objections to experiments in education are similar to those raised against their use in medicine. For example, the ethical issues concerning which classrooms get the new program and which ones wait until the following year echo the concern in medicine but, in comparison, can be more readily overcome in education. However, the place where the parallel breaks down, and where we will focus the remainder of this article is the practical matter of getting the same program to happen in multiple classrooms. Implementing a consistent treatment is far more problematic in education than in medicine. Administering a medication rather than a placebo is a fairly trivial procedure to replicate exactly on a large scale. And even with a complex new surgical procedure, one can assume that the surgeon is well trained in the procedure and all possible precautions are taken so as to avoid killing the patient in the process. Getting an education program to happen reliably can present a major challenge to the application of experimental research methods.

The problem of replicating the "treatment" in education

Before we can consider subjecting the "treatment" or "intervention" (e.g., an educational program, textbook, software, approach) to trials with randomized assignment of "subjects" (sometimes individual students but more often classrooms, schools, or districts), we must first find a way to make the same set of interactive tasks occur reliably in multiple settings. The fundamental concept in learning and development, that cognitive processes begin externally and especially in social interactions with adults and peers (e.g., working with manipulatives in a math lesson) before becoming abstract and internal (a symbolic cognitive procedure), entails that implementing educational programs requires establishing systems of interaction. A new educational program can never be a magic pill.

Teacher training or staff development are always major issues in replicating a new program. It is widely acknowledged that programs can fail for lack of sufficient training. Often an educational intervention will consist largely of staff development so the training is often itself the major challenge in implementation (e.g., if the new program calls for a different way of interacting with students, stronger subject matter knowledge, or new assessment procedures). For example, the introduction of computers into classrooms where the software is designed to be used as part of the classroom instruction or integrated with the curriculum can create massive variability because of differences in teacher training and in teacher interest in using computers. Unless the introduction of such software is combined with thorough training and commitment on the part of teachers, it is very difficult to show any effect. Even where training is provided, unless the training is done consistently, variability is introduced that drowns out the positive results from cases where the teachers were trained properly.

Eliminating differences among teachers to the greatest extent possible (not just controlling differences by randomly assigning them) is one family of strategies for effective experimentation. For example, by scripting the teacher-student interactions and rigidly enforcing compliance to the written script, the program eliminates most differences among teachers in their perceptiveness or skill in building scaffolding. A similar strategy is to focus on computer tutoring systems or micro-worlds in which children can participate with little or no support or supervision by teachers. Since the software will respond more or less consistently regardless of what brand of computer it is installed on, variability is again greatly reduced. As the perceptual

capability of computers improves and our understanding of teacher-student interactions grows, such systems become feasible.

Variability can be introduced from an enormous number of sources besides differences in the quality of training. From whatever the source, the greater the variability—even if randomly distributed between the treatment and a control group—the harder it is to see via statistical analysis whether there is any actual difference between the groups.

Lowering this variability is critical for successful experimentation. It allows researchers confidently to assert a beneficial effect of the program on the basis of a smaller numbers of subjects, which reduces the cost of the experiment, or on the basis of smaller actual differences between the treatment and control groups. The effort to control the treatment in schools, however, can run into the same issues of ecological validity that have long been a critique of psychological research in the laboratory.

An Example of Ecological Invalidity in Field Research

We introduced the concept of ecological validity in the context of psychological research in the laboratory. The idea is that the constraints, controls, and simplifications required for replication of the task in the experimental environment can result in a mismatch between the experimental treatment and how the task is approached when those controls are not in place. The same concept can also be applied to field research that uses experimental controls. When complex educational programs are implemented on a scale large enough to show an effect, the effort of maintaining tight control over the implementation has the danger of introducing ecological invalidity. If the researcher doesn't consider (a) the socio-ecological importance of teacher career paths, (b) the influence of site-based management, (c) the power of the local school board in making fiscal decisions, or (d) the constitutional role of states in setting educational standards, rigorous research can come to invalid conclusions about the applicability of its findings to educational practice.

An example of the potential for ecological invalidity is the case of a systematic attempt to enforce fidelity to a specific implementation across different educational jurisdictions that have been recruited to participate in an experimental effectiveness study. In the normal course of events, there is no requirement for one school district to implement an educational program (for example an intervention purchased from a publisher) exactly like its neighboring district let alone exactly like one in another state. Getting this additional effort to happen must be paid for by the researchers and their funding agencies and requires professionals in different school districts to follow explicit instructions from an individual (the researcher) outside the normal chain of command. There will now be a problem in generalizing the experimental findings to the case of school districts that are no longer required to follow the script after purchasing the intervention. The ecology of the real implementation is different from that which enforced compliance in the experiment.

A school district administrator considering purchasing the program may be persuaded by the research report stating that unless the district enforces fidelity to the experimental implementation, positive results will not occur. But unless the experimental program tried multiple variations on the implementation and did so across many different populations (an activity that would require ever stronger controls to assure that the distinctions between implementations were precisely maintained), the administrator actually has no evidence that the publisher had invented the best possible implementation of this program. The administrator could not be certain that his or her own staff might not find an even more effective way of implementing the program. In so far as the requirement for experimental control shuts down the potential for local improvements in the program, the experimental research may be trading off higher levels of overall effectiveness simply for lower variability.

A similar ecological issue arises if the administrator asks whether the massive effort needed simply to replicate the "treatment" may itself have beneficial side effects that have little to do with what the developers of the program had in mind. Did the training itself or greater time-on-task in classrooms (regardless of the particular materials) really make the difference? Researchers may be better advised to allow local variations so as to

discover best practices empirically. It may be better to invite local innovations that can improve the practices on the basis of the practitioners' own observations of what works for their students and their staff. This approach will make it more difficult to demonstrate the effectiveness of a narrowly prescribed program but may result in local experiments that are valid within the local socio-ecology.

Conclusion

The effort to apply large-scale effectiveness trials to education opens up a new gap between theory and practice. Because of the difficulty of replicating the same classroom events on a large scale, researchers have to take several steps back from the actual classroom practice, which then merges with the statistical noise. This is not to say that the desire to use empirical evidence to improve educational outcomes is misplaced, or that random assignment of units to conditions is somehow misconceived. But certainly problems result from the enormous difficulty in replicating a treatment in education, which puts it far removed from replicating a medical treatment.

It is possible to bridge this gap but it requires going back to what we've learned about the nature of teaching, learning, and development. First, there is wide consensus among education scientists that children actively try to learn and try to make sense of what they hear as well as the tools they are given. Likewise, teachers are continually constructing environments for children to participate in. Even when the teaching style is rote recitation, there is a dialog in which children participate, albeit in a highly structured manner. The children's response is information for the teacher and the teacher's response is information for the children. If we think of teaching and learning as a very local information system, we can begin conceiving of educational programs as things that teachers and children do, not things that are done to them.

The vendors of large complex programs such as textbooks or district-wide interventions will prefer to have them proven such that the successful results of a single study can be claimed to predict success in any school district. The federal legislation, however, makes it incumbent on their grant recipients, the schools and districts, to prove the effectiveness of their programs. The school does not have to show that its program will work in a neighboring district; it only has to show that it is likely to work again next year with the friends and relatives of the current cohort. Effectiveness evidence gathered on a national scale will, in fact, have less direct applicability than evidence collected by a particular school district when it comes to a purchasing decision. Such local experiments, by and for a local district, are becoming feasible as electronic information systems allow schools and districts to analyze locally generated data.

Pilot experiments in a school district that make use of local data can be part of an information system that helps teachers see what works and increases their interest and commitment, whether the program is something homegrown or purchased from a commercial vendor. Likewise, giving children immediate feedback on what works and what to do to make more progress should also be useful in capturing their attention and motivation to learn.

What is the role for education researchers? As a scientific community we need a better understanding of how learning and teaching work in actual schools. Conducting large-scale replications of programs is a technical task that ultimately has little to do with gaining an understanding of causal processes. Given the gap between the laboratory and the classroom, smaller-scale formative research in classrooms, as well as qualitative analyses of teacher-student interactions, can be very effective in identifying the reasons that educational practices work or don't work.

Since our initial work together on the gap between the laboratory and the classroom, we have taken considerably different paths. Cole has made it a firm practice never to tell teachers what they should do on the basis of his research (and to be wary of the well-intended advice of many others). He has, instead, focused on teaching and learning that occurs outside of school in informal settings where children participate voluntarily. Teachers are always welcome in such settings, and often they are motivated to change their practices by what they observe. Newman has spent considerable time designing and testing software for learning and communication in schools, including most recently,

systems to support research by school districts. While the goal has been software that can have widespread utility and commercial success, the development process has always come back to small-scale formative tests with practitioners. In both cases, implementation of theory in practice has become the practice, a mode of work we find theoretically as well as practically fruitful.

References

Bransford, J.D., Brown, A.L., & Cocking, R.R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, *50,* 255-272.

Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Harvard University Press.

Cole, M., Hood, L., & McDermott, R. (1979). *Ecological niche picking: Ecological invalidity as an axiom of experimental cognitive psychology*. New York: Laboratory of Comparative Human Cognition, The Rockefeller University.

Cook, T.D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

Erickson, F., & Gutierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher, 31*(8), 21-24.

Gibson, J.J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Jacob, E., & White, C.S. (Eds.). (2002). Theme issue on scientific research in education. *Educational Researcher, 31*(8).

Lewin, K. (1943). Defining the "field at a given time." *Psychological Review*, *50,* 292-310.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Neisser, U. (1976). *Cognition and reality*. San Francisco: W.H. Freeman.

Newman, D., Griffin, P., & Cole, M. (1989). *The construction zone: Working for cognitive change in school.* Cambridge, MA: Cambridge University Press.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: W.W. Norton.

Raudenbush, S. (2002, February). *Scientifically based research.* Paper presented at a seminar sponsored by the U.S. Department of Education, Washington, DC.

Rogoff, B. (2003). *The cultural nature of human development*. New York: Oxford University Press.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wertsch, J.V. (1986). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Whitehurst, G.J. (2002, February 28). [Statement of Grover J. Whitehurst before the Subcommittee on Education Reform Committee on Education and the Workforce, U.S. House of Representatives]. Washington, DC.