Classroom Research and Cargo Cults

By E.D. Hirsch Jr.

> "We really ought to look into theories that don't work, and science that isn't science. I think the educational . . . studies I mentioned are examples of what I would like to call cargo cult science. In the South Seas there is a cargo cult of people. During the war they saw airplanes with lots of good materials, and they want the same thing to happen now. So they've arranged to make things like runways, to put fires along the sides of the runways, to make a wooden hut for a man to sit in, with two wooden pieces on his head for headphones and bars of bamboo sticking out like antennas — he's the controller — and they wait for the airplanes to land. They're doing everything right. The form is perfect. It looks exactly the way it looked before. But it doesn't work. No airplanes land. So I call these things cargo cult science, because they follow all the apparent precepts and forms of scientific investigation, but they're missing something essential, because the planes don't land."
>
> — Richard P. Feynman, "Cargo Cult Science," *Surely You're Joking, Mr Feynman!: Adventures of a Curious Character* (Norton, 1985).

After many years of educational research, it is disconcerting — and also deeply significant — that we have little dependable research guidance for school policy. We have useful statistics in the form of test scores that indicate the achievement level of children, schools, and districts. But we do not have causal analyses of these data that could reliably lead to significant improvement. Richard Feynman, in his comment on

---

E.D. Hirsch Jr., a distinguished visiting fellow at the Hoover Institution, is president of the Core Knowledge Foundation and professor emeritus of education and humanities at the University of Virginia. *Special thanks to Eric Hanushek, Liz McPike, Ralph Raimi, Louisa Spencer, Steven Stahl, and Grover Whitehurst for their suggestions.*

"cargo cult science," identifies part of the reason for this shortcoming — that while educational research sometimes adopts the outward form of science, it does not burrow to its essence. For Feynman, the essence of good science is doing whatever is necessary to get to reliable and usable knowledge — a goal not necessarily achieved by merely following the external forms of a "method."

The statistical methods of educational research have become highly sophisticated. But the quality of the statistical analysis is much higher than its practical utility. Despite the high claims being made for statistical techniques like regression analysis, or experimental techniques like random assignment of students into experimental and control groups, classroom-based research (as contrasted with laboratory research) has not been able to rid itself of uncontrolled influences called "noise" that have made it impossible to tease out the relative contributions of the various factors that have led to "statistically significant" results. This is a chief reason for the unreliability and fruitlessness of current classroom research. An uncertainty principle subsists at its heart. As a consequence, every partisan in the education wars is able to utter the words "research has shown" in support of almost any position. Thus "research" is invoked as a rhetorical weapon — its main current use.

In this essay I shall outline some fundamental reasons why educational research has not provided dependable guidance for policy, and suggest how to repair what it lacks. On a positive note, there already exists some reliable research on which educational policy could and should be based, found mainly (though not exclusively) in cognitive psychology. In the end, both naturalistic research and laboratory research in education have a duty to accompany their findings with plausible accounts of their actual implications for policy — as regards both the relative cost of the policy in money and time and the relative gain that may be expected from it in comparison with rival policies. Including this neglected dimension might wonderfully concentrate the research mind, and lead to better science in the high sense defined by Feynman.

A tale of two studies

THE NOVEMBER 2001 issue of *Scientific American* includes an article called "Does Class Size Matter?" about the policy consequences of research into the beneficial effects of smaller class size. The centerpiece of the article is the famous multimillion dollar STAR (Student/Teacher Achievement Ratio) study — considered to be a methodological model for educational research — which showed with exemplary technique that reducing class size will enhance equity and achievement in early grades.

But when California legislators dutifully spent $5 billion to reduce class

size in early grades, the predicted significant effect did not result. Educational researchers, including the authors of the *Scientific American* article itself, complained that the California policy was implemented with "too little forethought and insight." Presumably this complaint implies that there are many factors that affect educational outcomes, and that we should not rely on a single one like class size. This after-the-fact criticism is valid. But if the California legislators had searched for useful "insight" in the STAR research they would have been disappointed. "Forethought and insight" cannot compensate for the deeper problem that *the process of generalizing directly from classroom research is inherently unreliable.*

Also in November 2001, there appeared an article in *Education Week* that summarized research into the multimillion dollar "whole-school reform" effort ("Whole School Projects Show Mixed Results"). According to the article, the researchers could not reliably discriminate between those programs that worked well and those that did not. The evaluators blamed the inconclusiveness of the results on uneven implementation of the various programs by the schools — an unhelpful observation. As a consequence, neither the expensive "whole-school" programs nor the expensive research into their effectiveness can usefully guide policy across the nation — which was a chief aim of the enterprise.

Neither the expensive whole-school programs nor the expensive research into their effectiveness can usefully guide policy.

These are but two recent examples of the general inconclusiveness of educational research. The historical record — like these two particular studies — supports Feynman's contention that even when educational research follows the external forms of science, it misses the essence. It dutifully gathers complex data, and uses control groups and experimental groups, and it applies sophisticated statistical techniques. In rare cases, as in the STAR study, it follows the still more rigorous practice of purely random assignment of students to the experimental and control groups. But even after researchers have dutifully followed "all the apparent precepts and forms of scientific investigation," the planes don't land. The test-score gaps between social classes do not narrow.

What is missing from this research? How, for example, might the STAR study have been made scientifically more solid, and ultimately more useful for the policymakers of California? These improvements would not have been achieved by using the now widely advocated technique of random assignment, since random assignment was in fact used. In fact, it was not the experimental structure of STAR but its intellectual structure that was deficient. This multimillion dollar study does not hazard a clear and detailed theoretical interpretation of its own findings. It does not, for example,

answer such nitty-gritty questions as: What are the various causal factors that make smaller class size more effective for earlier grades than for later ones? Could there be alternative and even more reliable ways of achieving similar or higher student gains? Much of the literature I have read in connection with STAR quietly assumes that smaller class size is itself the causal agent. But even the more sophisticated interpretations of STAR which posit deeper causal factors do not systematically explore the following critical issue: Given the probable causes of student gain, are there even more effective and less costly ways of applying those causes and achieving the same or greater gains? If, for example, an important causal advantage of smaller class size is more interaction time between student and teacher, are there alternative, less expensive policies for achieving more interaction time and even greater student gains? These are the questions that a policymaker needs to have answered, and it is the duty of the informed researcher on the ground — not the beset legislator — to ponder and answer those questions.

Traditionally, scientific work is considered "good" if its results foster deeper theoretical understanding. One of the most disdainful remarks in the sciences is that a piece of work is "a-theoretical." It's true that in common parlance the word "theory" has an overtone of impracticality. Scientists, however, regard the formulation of theories about deep causal factors to be the motive force of scientific progress — a view that has rightly replaced an earlier just-the-facts conception of scientific advance. The STAR study is a first-rate illustration of the way in which the a-theoretical tradition in education research hinders its utility. Wolfgang Pauli once remarked about a scientific paper: "It is not even wrong." That is exactly what can be said about the STAR study, and by extension many other classroom studies. Most of them are profoundly a-theoretical. They neither enable good policy inferences nor advance the research agenda. And they have other problems as well.

Difficult and undependable research

AN APOLOGETIC ARGUMENT heard in education schools is that educational research can never be as clean and decisive as controlled laboratory experiments because, on ethical grounds, one cannot treat children like rats in a maze. Admittedly, there is truth in this defense. Even the most carefully conducted school research must operate in circumstances that preclude certainty. Unfortunately, however, the limitations of classroom research eliminate not only certainty, but also the very possibility of scientific consensus — a very serious problem indeed.

If we take an example of the best educational research — say the Tennessee class-size experiment — and ponder why it fails to serve policymakers well, some very basic reasons present themselves. The STAR researchers were at pains not to interfere with anything in the school setting

except class size. Had they manipulated other factors, they would have introduced unmanageable uncertainties into the analysis. They wanted to disclose what might be expected if the only policy change was the reduction of average class size from 24 to 15. Given such careful control and analysis, why was class-size reduction so much less effective in California than it seemed to be in Tennessee? There's one immediate and self-evident answer: In some settings, class-size reduction helps an average .2 of a standard deviation; in other settings it helps only .075 of a standard deviation (neither effect being much to write home about).

This simple restating of the results, while almost too obvious to mention, goes straight to the heart of one educational-research problem: the fact that results cannot be generalized. Such research carries with it an implicit claim to reproducibility in other settings. Otherwise, why undertake it? But its multiplex character almost guarantees non-reproducibility. If just one factor such as class size is being analyzed, then its relative contribution to student outcomes (which might be co-dependent on many other real-world factors) may not be revealed by even the most careful analysis. On the other hand, if other classroom factors had been experimentally controlled at the same time, then it would be extremely hard if not impossible to determine — even by the most sophisticated means — just which of the experimental interventions caused or failed to cause which improvements. And if a whole host of factors are simultaneously evaluated as in "whole-school reform," it is not just difficult but, despite the claims made for regression analysis, impossible to determine relative causality with confidence.

In his essay on cargo cult science, Feynman described how one researcher managed with great persistence finally to obtain a reliable result in studying rats in a maze. Here is his description:

> There have been many experiments running rats through all kinds of mazes, and so on — with little clear result. But in 1937 a man named Young did a very interesting one. He had a long corridor with doors all along one side where the rats came in, and doors along the other side where the food was. He wanted to see if he could train the rats to go in at the third door down from wherever he started them off. No. The rats went immediately to the door where the food had been the time before.
>
> The question was, how did the rats know, because the corridor was so beautifully built and so uniform, that this was the same door as before? Obviously there was something about the door that was different from the other doors. So he painted the doors very carefully, arranging the textures on the faces of the doors exactly the same. Still the rats could tell. Then he thought maybe the rats were smelling the food, so he used chemicals to change the smell after each run. Still the rats could tell. Then he realized the rats might be able to tell by seeing the lights and the arrangement in the laboratory like any commonsense person. So he covered the corridor, and still the rats could tell. He finally found that they

> could tell by the way the floor sounded when they ran over it. And he could only fix that by putting his corridor in sand. So he covered one after another of all possible clues and finally was able to fool the rats so that they had to learn to go in the third door. If he relaxed any of his conditions, the rats could tell.

As mentioned, given ethical constraints, the likelihood of conducting such a scientifically rigorous experiment on American schoolchildren would appear to be rather low.

There are other fundamental difficulties standing in the way of generalization from classroom research. Young children learn slowly. The cumulative effects of interventions are gradual, extending over years. Yet most educational research is conducted over spans measured in months rather than years, ensuring that effect sizes will tend to be small. These effects may be rendered almost invisible by another difficulty — the fact that the process of schooling is exceedingly context-dependent. Children's learning is deeply social, lending each classroom context a different dynamic. Moreover, learning is critically dependent on students' relevant prior knowledge. Neither of these contextual variables, the social and the cognitive, can be experimentally controlled in real-world classroom settings. The social context of schooling depends on unpredictable interactions between teachers and students, and among students themselves. And what students bring to a classroom depends not only on what they previously learned in school, but also — as is well-established — on unpredictable knowledge they gained outside of school.

Children learn slowly, yet most educational research is conducted over spans measured in months rather than years.

Detailed analyses of the contextual factors that influence learning are greatly to be desired, of course, but progress in understanding those contextual factors is unlikely to result from coarse-grained classroom studies. Progress is more likely to result from highly controlled "artificial" experiments that reveal the fine-grained underlying causes. It used to be thought that damp, low-lying air causes "swamp fever." (The other term for swamp fever, "malaria," means "bad air.") That theory of the cause of the disease was accepted by medical science as long as researchers stuck to coarse-grained observations which indicated that if you live in a swamp you are likely to get swamp fever from the bad air. It was not until the disease was put under the microscope that progress began to be made in determining the true causes and vectors of malaria. Medical science continues to advance as it becomes allied with ever more refined laboratory understandings. Its most striking and reliable advances have occurred since medicine became closely

tied to biochemistry at a still more fine-grained level — the molecular. By analogy, it is plausible to think that progress in educational research, if it occurs at all, will follow this sort of pattern.

Another hard-to-control contextual variable is, of course, teacher quality. One argument of this essay is that deep-lying principles of learning are more reliable than specific teaching methods, because a decision about which teaching methods will be most effective will depend on unpredictable contextual variables, with the result that the same underlying principle may require very different methods in different contexts. This means that the teacher's role as the on-the-spot translator of principles into methods is critical. But teacher training, though crucial, is not my subject here. Leaving aside the vexed and critical question of "teacher quality," the two other uncontrolled-for context variables that I mentioned — the social and the cognitive — are so important that their influences alone tend to drown out most experimental interventions. That will be true even when (as in STAR) the number of students being sampled is large enough to allow the hopeful assumption that the variables will cancel out. In those cases, the influence of contextual variables has been so great that the effect-sizes of most experimental interventions have been small.

The smallness of effect sizes has prompted disinterested scholars like H.J. Walberg, Barak Rosenshine, and Jeanne Chall to analyze whole masses of relevant studies on given educational topics to see if a reliable pattern emerges. These meta-studies are the most dependable sources of the meager insights that educational research has uncovered. But the end result of these painstaking analyses is that most conclusions still remain insecure, and still reflect the uncertainty and ambiguity of the underlying studies.

To summarize so far: Educational data are difficult to apply in a dependable way because of contextual variables that change from classroom to classroom and from year to year, and that drown out the effects of single or multiple interventions. Clearly, therefore, one major assumption of educational research needs to be examined and modified — i.e., the assumption that data about what works in schools could be gathered from schools and then applied directly to improve schools.

Changing the thought model

IS THERE A WAY in which this inherent uncertainty principle in educational data can be diminished? Yes, by placing less reliance on traditional educational research that makes inferences from school data and applies those inferences directly back to schools.

Here is an example of traditional educational research in action from the government's educational database called ERIC:

ERIC NO: ED394125 TITLE: Vocabulary Teaching Strategies: Effects

on Vocabulary Recognition and Comprehension at the First Grade Level. AUTHOR: Peitz, Patricia; Vena, Patricia PUBLICATION DATE: 1996

ABSTRACT: A study examined teaching methods for vocabulary at the first grade level. The study compared teaching vocabulary in context and teaching vocabulary in isolation. Subjects were 32 culturally diverse first-grade students from varying socio-economic backgrounds. The sample consisted of 14 boys and 18 girls, heterogeneously grouped. Two teacher-made tests were used, each consisting of 30 multiple choice items: Test A, to test vocabulary in isolation; and Test B, to test vocabulary in context. Target words for the tests were taken from the Dolch list, the Harris-Jacob list, and the reading material used in the classroom on a regular basis. Both tests were administered as pretests prior to instruction. After a 3-month period of instruction, Tests A and B were readministered as posttests to determine students' vocabulary growth. Results indicated that there was no significant difference in vocabulary acquisition by the sample. Results also indicated that, although there was vocabulary growth with both methods, the sample group's growth in vocabulary taught in isolation was greater than that of the vocabulary taught in context. Findings suggest that both methods of learning vocabulary will enable children to increase their vocabulary base and should be used. (Four tables of data are included; contains 37 references, 4 appendixes containing lists of vocabulary in context and in isolation, and related literature on vocabulary building.)

To paraphrase, there seems to be a slight benefit to teaching high-frequency vocabulary words in isolation rather than in context, but no significant difference in vocabulary growth as between the two methods. If the experiment had been made on a grander scale with thousands of students, random assignments, and a duration longer than three months, the data might have shifted in favor of teaching words in context. To repeat, however, it is unlikely that the results of a more massive experiment would supply dependable guidance. Again, we simply do not know enough about the uncontrolled factors at play in either sort of result to move confidently from research to policy.

But suppose a policymaker had to form a decision on how teachers should best achieve first-grade vocabulary enhancement (an extremely important issue). What decision should be made? Someone who read the work of cognitive scientists (rather than classroom reports) would find well-tested advice on how to teach vocabulary. They would find a consensus that, depending on the prior knowledge of students, both isolated and contextual methods need to be used — isolated instruction for certain high-frequency words students may not know or may not recognize by sight, like the prepositions "about," "under," "before," "behind," but carefully guided contextual instruction for other words. Teachers and administrators would learn

that word meanings are acquired gradually over time through multiple exposures to whole systems of related words, and that the most effective type of contextual word study is an extended exposure to coherent subject matters.

This scientific consensus arose not just from classroom educational research but principally from laboratory studies and theoretical considerations unconnected to the classroom. One theoretical consideration, for instance, is that a top-of-the-class 17-year-old high-school graduate knows around 60,000 different words. That averages out to a learning rate of 11 new words a day from age two. Although this estimate varies in the literature from 8 to 18, its range implies by any reckoning a word-acquisition rate that cannot be achieved by studying words in isolation. There is notable cognitive research on the subject of vocabulary acquisition. Synthesis of this research is a more dependable guide to education policy than the data derived from classrooms.

The most reliable guidance is not found in data from schools but rather in inferences from the laboratory.

If we follow this line of thought where it leads, we come to the conclusion that the most reliable guidance to what works in school is not to be found by looking at data from schools but rather by looking at inferences from the laboratory. ("By indirections find directions out.") Of course, these scientific inferences must prove themselves in the schools; they can't be permitted to produce worse educational outcomes than we had before. But because of the variability of the local contexts from which the school data is taken, the probability that an inference from school data is wrong is much greater than the probability that a scientific consensus is wrong.

Education-school proponents of "qualitative" research criticize quantitative research by taking note of the variability of classroom contexts, and claiming that all education, like all politics, is local. (They use the term "situated learning.") They pride themselves on following "ethnographic" methods, and taking into account the uniqueness of the classroom context. They rightly object that quantitative research tries to apply oranges to apples. But if their descriptions do not disclose something general that I could confidently apply to my own classroom, their studies are not very useful. And if their inferences did have general application, then the value of an "ethnographic" rather than a straightforward general description would lie in the literary vividness of a concrete example. But literary value is rarely claimed or observed in these productions.

Descriptive educational research suffers a fundamental shortcoming. To describe is to select what is important to describe out of an uncountable multitude of classroom happenings. How do I know that the chosen events are the ones that have made a difference? Overt behaviors like calling on shy

students or building medieval castles out of milk cartons may or may not be the behaviors that have mainly caused one classroom to learn more about medieval castles than another. To be useful, even in the abstract, the descriptions would have to be selected on the basis of a prior theory about what is important to be described. This begs the research question. What is important to be described is what careful research should be trying to find out, not what it should be taking for granted. Although advocates of qualitative research are right to point out the unreliability of quantitative analyses like the STAR study, they need to apply a similar skepticism to their own efforts.

The reliability picture changes dramatically when we apply consensus science to education. Cognitive scientists have reached agreement, for example, about the chief ways in which vocabulary is acquired. This theory gained consensus because it explains data from many kinds of studies and a diversity of sources. While incomplete in causal detail, it explains more of what we know about vocabulary acquisition than does any other theory. When we apply it, we are no longer applying oranges to apples, but well-validated general principles to particular instances, in confidence that the principles will work when accommodated to the classroom or other context.

A teacher needs not just practical maxims but also underlying general principles.

One might object that teachers should not have to think back to first principles every time they make lesson plans. Highly probable maxims that work most of the time (Francis Bacon called them "middle axioms") get us through the day. True enough, but for reasons I have already advanced, classroom research has been undependable in offering middle-level generalizations. Its maxims tend to be overgeneralized beyond their highly uncertain sphere of validity, so they are often inapplicable to particular circumstances. Teachers who were to read a different research report such as ERIC ED246392 or ED392012 would conclude they should favor the words-in-context approach.

Yet neither conclusion would be warranted. According to more general principles gleaned from cognitive science, it would be premature for teachers to follow either approach without further consideration. If students in a particular class already know and recognize by sight critical foundational words like "under," "over," "about," "beside," beneath," it wastes class time chiefly to use a words-in-isolation approach. This more general maxim is grounded not just in classroom research but in an interpretation of data from a diversity of domains.

Middle axioms are inherently probabilistic, and, in education, the probabilities change greatly in different circumstances. A teacher needs not just practical maxims but also underlying general principles that can guide their intelligent application. The wider public shows an understanding of this

truth in the adage "teaching is an art, not a science." This is another way of saying that the variabilities of classrooms demand a flexible application of deep general principles, not a mechanical application of methods and maxims.

What are "reliable general principles"?

FIFTY YEARS AGO, psychology was dominated by the guru principle. One declared an allegiance to B.F. Skinner and behaviorism, or to Piaget and stage theory, or to Vygotsky and social theory. Today, by contrast, a new generation of "cognitive scientists," while duly respectful of these important figures, have leavened their insights with further evidence (not least, thanks to new technology), and have been able to take a less speculative and guru-dominated approach. This is not to suggest that psychology has now reached the maturity and consensus level of solid-state physics. But it is now more reliable than it was, say, in the Thorndike era with its endless debates over "transfer of training."

To lend some credence to the proposition that general cognitive principles tend to be more dependable than maxims from direct classroom research, I shall now outline some issues in cognitive science about which a degree of consensus has been reached. Shrewd applications of these consensus principles would almost certainly enhance classroom learning, and ought also to encourage a shift in the way policymakers use educational data and research.

Prior knowledge as a prerequisite to effective learning. I have put this principle first, because so many other principles and policy implications flow from it. If "fortune favors the prepared mind," so does learning. One of the themes currently dominant in our education schools is that learning should be based on the mastery of formal habits of thinking rather than on "mere facts," that learning how to think is more important than mere accumulation of "factoids." The modicum of truth in this widely-held notion would appear to go something like this: After a student has reached a certain threshold of enabling knowledge, then acquiring a habit of critical thinking may be more valuable than acquiring a few more facts.

But it would be a profound mistake, uncountenanced by cognitive science, to suppose that skillful thinking can be mastered independently of broad subject-matter knowledge. The fallacy of derogating content is obvious in mathematics, where everyone concedes that skill and understanding in multiplication depend on a preparatory knowledge of addition. And the principle of preparatory knowledge applies not just to math, but to most other intellectual domains.

The research that offers the most dramatic evidence that relevant prior knowledge is critical to thinking skill is the area of expert-novice studies. The expert learns more from a given experience than a novice does, even

though the novice has much more still to learn. That's because being presented with too many not-yet-interpreted items overloads and confuses the mind, whereas prior knowledge makes experience salient and meaningful (see "meaningfulness" below), and the expert need interpret less novelty than the non-expert (see "attention" below).

Meaningfulness. A lot of learning is, of necessity, pretty meaningless. The connection between the sound and the sense of many words is entirely arbitrary. That the words "brother" or "sister" sound like they do is, for a child, just a brute fact that has to be learned. But, once the arbitrary sound-sense connection is learned, the meaningfulness of those words ensures that they will be remembered. Meaningfulness implies connectedness by experiential association (episodic memory), by schematic structure (semantic memory), or by emotional associations. In the expert-novice experiments, it is thought that prior knowledge enables the expert not only to connect the elements of an experience, but also to pick out what is meaningful and salient in it. Moreover, prior knowledge enables the expert to deduce more from the experience than the novice can. A novice looking at the outside of an Italian villa wouldn't understand that it has an unseen central courtyard; the expert, equipped with prior knowledge, would comprehend the unseen interior courtyard as well as the exterior walls.

One of the tasks of teaching is to make learning meaningful.

The familiar distinction between "rote learning" and "meaningful learning" is thus well grounded — if understood liberally. But, since not all learning is inherently meaningful to a child (e.g. "SIS-tuh," "BRUH-thuh") one of the tasks of teaching is to make it so. A brilliant kindergarten teacher once described to me some tricks she used to teach children the names of the numbers. One trick was to bring in a pretzel, "Look, this is the shape of the number 8." She plopped it into her mouth. "Look, I ate it! I 8 it." It's hard to believe that this method of making "rote learning" meaningful, which incidentally invoked the children's prior knowledge of the verb "to eat," could have been easily forgotten by the children.

The right mix of generalization and example. Learning in school requires generalization. Nothing could be more abstract and general than arithmetic. But to acquire the concepts of addition, subtraction, multiplication, and division (or as Lewis Carroll would have it: "Ambition, Distraction, Uglification, and Derision"), you have to learn more than the abstract conceptions. You have to work with a lot of examples. No one advocates saying to first graders "OK, kids, this is the commutative law of addition. You memorize that — and never mind fiddling around with all those beans." The beans or their equivalent are absolutely essential.

The optimal mode for learning most subjects is through a carefully devised combination of the general concept and well-selected examples. This

idea of teaching by both precept and example is so old — going back to the earliest literature in many cultures — that its confirmation in experiment is no surprise. Examples serve a number of functions that can't be retailed here. Researchers say that it's important to get the right mix and number of examples. If arithmetic exercises are too numerous and similar, time will be wasted. It is important to vary the angle of attack in examples, to illustrate different key aspects of the underlying concepts, and not to forget that explicit restatements of the general concept are equally important. The way we store these concepts is typically enmeshed with models or examples. One famous experiment showed that the concept "bird" is stored (by North Americans) as something about the size of a robin, not the size of a hummingbird or ostrich. Concept and example are deeply connected with one another in how we think and remember as well as how we learn.

Education aims to enable the mind to transcend the narrow constraints of working memory.

Attention determines learning. Although "motivation" and "interest" are perennial themes of education, and important to any practicing teacher, it is sobering to discover from cognitive science that motivation is only an indirect and dispensable aid to learning. Intention to learn, whether internally imposed by intrinsic interest and ambition, or imposed from outside through rewards and punishments, may be sometimes a condition for learning, but it is not a necessary or sufficient condition. Some things that we involuntarily pay attention to are learned and remembered better than things we are trying to learn and remember. What is learned is that which is paid attention to, and, typically, what is paid attention to is what is learned.

Attention is an aspect of our "working memory," a function that lasts just a few seconds. Out of the whirr of perceptual features that impinge on working memory every instant, we attend only to a salient few. That few is very, very limited in number, even for the most brilliant minds. A famous article about the limited number of things we can attend to at one time was called "The Magical Number Seven, Plus or Minus Two" (by G.A. Miller, first published in 1956). In some cases the limiting number is nearer to four. An expert with prior knowledge will be able to attend to many more things than a novice, not because of greater mental capacity, but because of "chunking." For an expert, noticing one thing is automatically to notice a myriad of things implied by it and known to be chunked with it, whereas the novice has to get through dozens of connections, which, because of the limitations of working memory, is impossible.

One chief aim of education is to enable the mind to transcend the narrow constraints of working memory by concentrating an immense wealth of individual elements into a single symbol or name that can be attended to all at

once. This concentration effect is one of the marvels of language, and it illustrates the immense importance of imparting a sufficient vocabulary. As individuals and societies learn more, they form and learn new names for these large complexes of concepts and perceptions. By means of effective names and symbols, the vastness of what an ordinary school child can retain, use, and pay attention to in, say, mathematics, exceeds the capacity of the most learned doctors of fourteenth-century Oxford.

If the attended-to things are given meaning by being connected with what we already know, we will learn (remember) them. If we do not attend to them and do not accommodate them to some known structure, we will usually not learn them. Although this finding is not surprising to common sense, it is a sobering reminder that we should not be overly distracted by the vast and unreliable literature on what will or will not properly motivate students — a debate that seems baffling to many teachers, since what motivates some students does not motivate others. A teacher's job is to ensure meaningful attention by as many students as possible towards that which is to be learned — using whatever methods may come to hand, including, above all, giving students the preparatory knowledge that will make attention meaningful.

Effective learning depends on rehearsal by one means or by another.

Rehearsal (repetition) is usually necessary for retention. How long something will be remembered is typically determined by how often it has been attended to. Rehearsal has the double purpose of retention and making meaningful connections between experiences. There is evidence that the need for rehearsal has a physical basis in the neuron structure of the brain. The need for repetition to maintain what is learned has been well understood in every culture. We teach children little poems or songs so they can retain the letters of the alphabet or the days of the months. All this the world knows well, however contemporary slogans may disparage it.

The disagreeable need for rehearsal is called in the educational parlance "drill and kill." Good teachers try to find ways of making rehearsal less obviously painful, when that is possible. But effective learning depends on rehearsal by one means or by another. In the old argument between "natural development" and "practice makes perfect," it is the latter that has the support of cognitive science.

Some useful findings can make practice effective. It has long been known that massed practice is less effective than distributed practice. Cramming for an exam is less effective for long-term retention than keeping up with assignments as you go along. Frequent classroom testing of students (another disparaged method) is a very effective distributed-practice technique. To test students shortly after they learn something rehearses that knowledge.

Moreover, students' awareness that a test is coming focuses their attention during original learning — giving classroom tests a double whammy for learning. A maximally effective mode of practice is to rehearse something just shortly before getting rusty, thus gradually extending the time span between rehearsals. So superior is distributed practice to massed practice that the cognitive scientist Ulrich Neisser was moved to poetry:

> *You can get a good deal from rehearsal*
> *If it just has the proper dispersal.*
> *You would just be an ass*
> *To do it en masse:*
> *Your remembering would turn out much worsal.*

Automaticity (through rehearsal) is essential to higher skills. Rehearsal serves other purposes beyond long-term retention and the constructing of meaningful connections. It also serves to make certain operations non-conscious and automatic. An obvious example is reading. The beginning reader must consciously correlate sound and symbol, and consciously move the eye from left to right, and consciously form the symbols into words. The beginning student does not have much "channel capacity" left for paying attention to what the words are saying. Since working memory can attend to just a few things at a time, the meaning of the sentence and even its component sounds are likely to spill into oblivion. As these underlying processes become more and more unconscious and automatic, the possibility grows for meaningful reading, and finally for thinking about the meaning. The processes do not become automatic just because children grow older, as the term "development" is often used to imply. Skills become automatic by being practiced.

What is true for reading is also true for other activities. Obvious examples come from sports; the more one has to think about all the motions required for hitting a tennis ball well, the less one is likely to do so. In sports no one doubts the need to gain automaticity. And it is no less true of other skills including academic ones. Automaticity frees up the working memory and allows it to concentrate on higher-order thinking.

Implicit instruction of beginners is usually less effective. A theme in the literature of American education research is that natural, real-world simulations (hands-on projects), in which the student gains knowledge implicitly, are superior to the artificial, step-by-step methods of traditional schooling. It is initially plausible that exposure to the complex realities of reading — the "whole language" method — would lead to more sophisticated reading skill than stumbling along step by step with the bricks and mortar of the alphabetic code. The more general question is this, however: Should students be immersed right away in complex situations that simulate real life — the method of implicit learning — or should they first be provided with explicit modes of instruction that are focused on small chunks deliberately isolated from the complexities of actual situations?

The answer one gets from cognitivists is complex. A teacher needs to

engage in both implicit and explicit teaching. Because of the limitations of working memory, a step-by-step, explicit approach is good for beginners. A new tennis player has to be able to hold the racket and hit the ball over the net, and usually needs instruction in those sub-skills before going on to play a game. On the other hand, it's hard to see how one could gain knowledge of the ways subskills work together except in an actual game. Successful coaches provide guided practice in isolated subskills, and also in how to put them together in real-world simulations.

Since the resolution of the implicit-explicit debate is that teachers should use both, the main point of considering the issue here is that explicit learning has been subjected to widespread "research-based" condemnation in education schools. Hence the subject forms a good illustration of the contrast between educational research and cognitive research.

Explicit learning has been subjected to widespread "research-based" condemnation.

There's a dramatic experiment in the literature. At issue was the problem of how to teach people to discern the sex of day-old chicks. The protosexual characteristics are extremely subtle and variable, and even after weeks of guidance from a mentor, trainees rarely attain a correctness rate of more than 80 per cent. Learning this skill has important financial implications for egg-producing farmers, and chick-sexing schools have been set up in Canada and California. The school training, which involves implicit learning from real-world live chicks, lasts from six to 12 weeks.

It occurred to two cognitive scientists familiar with the literature on implicit vs. explicit learning that these chick-sexing schools might present an experimental opportunity. They wondered if they could construct a more efficient learning program based on their knowledge of the literature. They decided to capitalize on the experience of a Mr. Carlson, who had spent 50 years sexing over 55 million chicks. From a set of 18 chick photographs representing the different types, Mr. Carlson was able to identify the range of critical features distinctive to each sex, and on the basis of his trait-analysis, a single-page instruction leaflet was created. Training was to consist in looking at this analytical leaflet for one minute.

To conduct the experiment, people without any chick-sexing experience were randomly divided into two groups, one of which looked at the leaflet. Thereafter, both groups were tested. Those who did not study the leaflet scored about 50 percent, that is, at the level of pure chance. Those who looked at the leaflet scored 84 percent, which was even better than the scores achieved by professional chick-sexers. Alan Baddeley, the distinguished psychologist from whose book this example was taken, interprets the experiment as "an extremely effective demonstration that . . . one minute of explicit learning can be more effective than a month of implicit learning."

Reading and other academic skills are, at least in some respects, analogous to chick-sexing. Mr. Carlson's 50 years of experience enabled him to isolate the protosexual traits of chicks into an analytical chart that could be learned in 60 seconds. This feat is analogous in its form to the achievement of ancient scholars in isolating the phonemic structure of speech into an alphabet of 26 letters. Their work, one of the great intellectual feats of human history, can now be recited or sung by a non-precocious preschooler by the age of four. Teachers and students can then be trained in the approximately 43 phonemes of English and their various correlations with the 26 alphabetic letters by using focused, analytical techniques. There is now ample evidence that carefully planned explicit instruction in phoneme-letter correlations is the fastest and surest way of empowering all beginners to decode alphabetic writing. In instances like these, explicit instruction with clearly defined goals is superior to implicit instruction and constitutes the most effective use of that precious commodity, school time.

Implicit rather than explicit learning is, as we have seen, the superior method for vocabulary growth, since word acquisition occurs over a very long period, and advances very, very gradually along a broad front. On the other hand, explicitly learning a few foundational words is much faster than implicitly learning them. It may be that explicit learning is best for a limited number of foundational elements, while implicit learning is best for advancing slowly on a broad front. It is not yet clear whether this division of labor between explicit and implicit learning applies to domains other than vocabulary growth, but even after that issue is sorted out, common sense will remain a valuable classroom commodity.

Of convergence and consensus

In recommending skepticism towards the findings of classroom research, I have at the same time counseled confidence in the findings of cognitive science as the more reliable guide to educational practice. Cognitive science, in contrast to school-based research, gathers data from many sources and explains why they converge on a consensus interpretation. I do not mean that cognitive research is always good or that educational research is always bad. The difference in the two fields is that, whereas classroom research, in the nature of the case, rarely converges on a consensus view, cognitive science has recently begun to do so.

The principle of independent convergence has always been the hallmark of dependable science. In the nineteenth century, for example, evidence from many directions converged on the germ theory of disease. Once policymakers accepted that consensus, hospital operating rooms, under penalty of being shut down, had to meet high standards of cleanliness. The case has been very different with schools. Educational policymakers, in the grip of their own strong sentiments or in thrall to the latest bulletins from the edu-

cation-research front, have authorized experimentation upon children on a vast scale, often under assumptions that conflict with the relevant scientific consensus.

What policymakers should demand from the research community is consensus. This has been achieved in some cases. Under the aegis of the National Institutes of Health, a high degree of consensus has been reached among both mainline psychologists and school-based researchers regarding effective modes of teaching early reading. This NIH work is notable for having integrated both laboratory and classroom research and for having supplied theoretical accounts of the underlying causal processes at a detailed level.

Without greater theoretical sophistication we are unlikely to achieve greater practical results.

Policymakers can further demand that laboratory researchers take the plunge that they have not yet taken and offer us theoretical extrapolations to classrooms. On the other side, policymakers can demand that classroom researchers take the extra effort and study needed to offer theoretical descriptions (deduced from laboratory research) of the causal factors that have produced the classroom results they report. This was the theoretical element so glaringly absent in the STAR study. The nation needs both groups — basic researchers and school-level researchers, acting in concert to begin a tradition of hard theoretical effort at the most profound and intricate level. Without greater theoretical sophistication we are unlikely to achieve greater practical results. With it, educational research could begin to earn the high gratitude and prestige that it currently lacks but which, given its potential importance, it could some day justify.

Recently, an impressive book on educational research has appeared called *Evidence Matters* (Brookings Institution, 2002). It contains a fine essay by Thomas D. Cook and Monique R. Payne advocating the method followed in the STAR study — random assignment of students into experimental and control groups. The Cook and Payne essay argues that randomization is the most convincing way to determine whether the outcomes of educational interventions have statistical significance. Currently, the method of random assignment is advocated as the herald of a new research era.

One may concede to Cook and Payne and others that the practice of random assignment may yield more convincing evidence of statistical significance than other methods of data gathering, but that is not to concede that statistical significance is itself a reliable guide to educational policy. When an intervention yields effects that have statistical significance, we can infer only that the effects are not accidental in the given circumstances. As was evident in the STAR study, we cannot necessarily be confident that the observed effect

size will be repeated in new circumstances.

Brute empirical data does not speak its own meaning. The main policy use of educational research is to enable us to make good predictions about which interventions will yield significant effects in new situations — by understanding of the root causes of the observed effects. In a domain as causally complex as mass education, "statistical significance" no matter how rigorously derived must be interpreted with a wary eye.

For instance, it is dangerous to predict long-term benefits from short-term results. Random assignment research has shown short-term gains from teaching "metacognitive" reading strategies (such as looking for the main idea). At the same time, cognitive theory predicts that the rate of student improvement with such interventions will not only reach a ceiling but will ultimately slow down a student's progress in reading — an important illustration that theory (based on extensive data) is more important and useful than ad hoc data.

In short while the new stress on random assignment is welcome, it doesn't affect the validity of Feynman's strictures about the limitations of method in educational research. A companion volume to *Evidence Matters* needs to be issued entitled *Theory Matters*. By all means let us use random assignments where plausible in educational data gathering. But then let us interpret the results warily in light of the deepest and most detailed theoretical insights into root causes that science has currently achieved.

In commenting on a draft of this essay, a federal administrator of research who has pursued both classroom and laboratory research observes that, ideally, the relationship between classroom research and cognitive science ought to parallel the collegial and fruitful relationship between medical research and biochemistry. This hopeful analogy, he concedes, could not be validly drawn in describing the educational research of the past, but he is determined to make the analogy more applicable in the future. Godspeed!