Random Thoughts on the (In)credibility of Educational–Psychological Intervention Research

Joel R. Levin

Department of Educational Psychology
University of Arizona

I am proud that throughout my professional career my path has crossed—if only symbolically—with that of the much-revered Professor Edward Lee Thorndike. For starters, Benton Underwood’s frequency theory, which years ago Elizabeth Ghatala and I used as a basis for several research investigations (e.g., Ghatala & Levin, 1976), has at least a psychological connection to Thorndike’s Law of Exercise. Then, while editing the 1992 American Psychological Association centennial issue of the Journal of Educational Psychology, I paid homage to Professor Thorndike’s enormous contributions to the field of educational psychology. In the inaugural issue of the Journal of Educational Psychology, Thorndike (1910) wrote on the potential contributions of education to psychology, and vice versa, thusly:

In all cases psychology, by its methods of measuring knowledge and skill, may suggest means to test and verify or refute the claims of any method. For instance, there has been a failure on the part of teachers to decide from their classroom experience whether it is better to teach the spelling of a pair of homonyms together or apart in time. But all that is required to decide the question for any given pair is for enough teachers to use both methods with enough different classes, keeping everything else except the method constant, and to measure the errors in spelling the words thereafter in the two cases. Psychology, which teaches us how to measure changes in human nature, teaches us how to decide just what the results of any method of teaching are. … [Yet,] not only do the laws derived by psychology from simple, specially designed arranged experiments help us to interpret and control mental action under the conditions of school-room life. School-room life itself is a vast laboratory in which are made thousands of experiments of the utmost interest to “pure” psychology. (pp. 7, 12; emphasis added)

I bounce off Thorndike’s now nearly 100-year-old thinking to introduce one of the major themes that I will touch on here while paraphrasing Simon and Garfunkel: “Where have you gone, Joe and Joan Educational Scientist? Our schoolrooms turn their lonely eyes to you!” Incidentally, to promote some sense of engagement in, and anticipation of, what is to come, I will also be citing famed educational philosophers Cuba Gooding, Jr., Aretha Franklin, and Robin Williams.

As a final Thorndikian instance, in an article that I collaborated on with American Psychological Association Division 15 colleague Angela O’Donnell (O’Donnell & Levin, 2001), we discussed the so-called wars that have been waged in educational psychology’s quest to find and define itself, between the epistemologies of E. L. Thorndike and another educational psychology pioneer and philosopher, John Dewey. On that topic, colleague Dave Berliner (1992) wrote:

It is sad, in a way, that two of the founding fathers of educational psychology fought for our soul, and only one of them won. Dewey, who was concerned about the systemic, the authentic, the ordinary thinking of teachers in schools, lost to Thorndike, the analytic, laboratory-based scientist, concerned with measurement, statistical inference, and the development of propositional knowledge. (p. 158)

Educational historian Ellen Lagemann (2002) concurs with that assessment. So with all due respect to Lagemann and Berliner, even—or especially—today, I would still select E. L. Thorndike as educational psychology’s number one draft choice. Consequently, I suspect that Thorndike would approve of at least some of my current reflections. And those reflections are directed primarily at the future livelihood of our discipline, aspiring educational researchers. In this article, I appropriate excerpts liberally and unashamedly from certain of my previous writings and, in particular, from Levin & O’Donnell (1999, 2000). I do so partly from an efficiency motive, but mainly because I am confident that few of the current readers have come across these obscure “classics” anyway. And so I begin with some random thoughts related to conducting what I refer to as scientifically credible research on educational interventions.

Requests for reprints should be sent to Joel R. Levin, Department of Educational Psychology, University of Arizona, Tucson, AZ 85721. E-mail: jrlevin@u.arizona.edu
Contrary to intuition, the concept of random is neither readily nor completely grasped by many, including the public, graduate students, and even seasoned researchers. Take the concept of random selection (or random sampling), for example.

Random Selection (or Random Sampling)

A doctoral candidate from an unnamed department at an unnamed university asked me to serve on his dissertation committee. As part of his study, one day he went out to a regional airport, where he executed his plan of interviewing every exiting passenger that he encountered on that particular day. When the student indicated that he had interviewed a random sample of airline passengers, I reflexively responded that his interviewees were neither random nor a sample.

Random sampling is necessary for generalizing from a sample to a population, an external validity construct from Campbell and Stanley (1963). Is random sampling always necessary and beneficial? That depends on the research purposes and questions. In very many research contexts, it is not just unnecessary; it is impossible (see, for example, Serlin, Wampold, & Levin, 2003). Yet, even in such contexts, it might not compromise the researcher’s conclusions one whit. In other situations, even when collecting a random sample is possible, actually collecting one might be ludicrous. A New York Times article on a proposed new system for improving the judging of international skating competitions represents a case in point. According to this system, the number of judges would be increased from the current 9 to 14—not a bad idea, given what we know about measurement characteristics. However, and this is where the random ludicrousness begins, from the population of 14 judges, the marks of only 7 randomly selected judges would figure into the actual scoring. What's wrong with this picture? Here’s how George Rossano (2002), author of the article, sees it (and very clearly, I might add):

Suppose … that you have black socks and blue socks in a drawer and you want to know which you have more of. … If you have seven black socks and seven blue socks in the drawer and [you] randomly pick out seven, you will always pick out more of one color than the other. You always get the wrong answer. Now try it with six black socks and eight blue ones. About one-third of the time, your random choice of seven will give a majority of black socks. Again the wrong answer—there are more blue socks in the drawer, not black. Go to the extreme and put 4 black socks in the drawer and 10 blue ones. Pick seven again several times. Sometimes you will still end up with a majority of black socks. Still the wrong answer. Over all, about one-quarter of the time you get the wrong answer for the mix of socks in the drawer by randomly picking 7 of 14. In a skating competition, one-quarter of the time the placements will be incorrect. The lesson is simple: If you want to know the number of socks in a drawer, nothing beats actually counting the socks, and if you want to know who should win a skating competition, nothing beats counting all the judges, preferably a large number of judges. (p. 7)

Brilliant! Of course, adjusting for judges’ known nationalistic biases is a different matter entirely and one that random sampling per se will not correct.

Now back to the scientific skating rink. Demographers and everyday pollsters employ some form of random sampling—simple, stratified, or cluster, to name a few—and they are well aware of the costs and benefits associated with each. The last of these has direct implications for its “random assignment” counterpart, which I next discuss.

Random Assignment (or Randomization)

Random assignment in a research context can be thought of alternatively as in one form or another assigning participants randomly to treatments or treatments to participants (with or without preassignment blocking constraints). Random assignment is a canon of scientifically credible research, an internal validity construct. I, for one, am solidly for randomization as well—even to the point of proselytizing, which I will do in abundance throughout the rest of this article.

As an example of where randomization, if properly understood, could have helped to avert a disaster, consider the 2000 Florida presidential election, the butterfly ballot, and the postmortem arguments about the randomness (read, unbiasedness) of voting machine error. Unfortunately, the errors cited, which include the flawed butterfly ballot, pregnant chads, and other mechanical impediments cannot be considered random. For example, in Palm Beach County, where voters encountered butterfly ballots, the candidates’ names appeared in a fixed order on the ballots. Thus, the claim that machine-scoring errors are random is not correct. Suppose, for sake of argument, that the mechanical vote-reading machine is able to detect a mark or punch at the top of the ballot more reliably than a mark or punch further down. Unless ballots are randomized with respect to the order of candidates’ names, machine error is systematic, not random. With respect to the two major contenders in question (Bush and Gore), this implies that to be considered an unbiased mechanical-counting procedure, half of all distributed ballots, randomly determined, would have had to list the names Bush and Gore in that order; and the other half, Gore and Bush in that order. Had that process been followed, then claims of machine random error (that is, unbiasedness) would be justified, even in the presence of other ballot booboos or mechanical misadventures. For instance, had systematic name randomization been incorporated, then the documented human voter error associated with the butterfly ballot would also have been truly random.

As long as we are in a balloting-and-assessment frame of mind, let me give you a little homework assignment, the an-
swept to which will be alluded to in various ways before this article has been completed. A common researcher misbelief is that in controlled intervention studies, random assignment to experimental conditions safeguards against virtually all criticism concerning the scientific credibility of the research. Comment on this statement in 50 words or less. Hint: Random assignment guarantees that the resulting groups that are formed are initially equivalent in all respects from a statistical probability standpoint.

But what precisely do these random thoughts have to do with the main thrust of my message concerning the plight of educational-intervention research today? Let me begin with the concept of evidence.

THE CONCEPT OF EVIDENCE

The popular media is responsible for bringing a good deal of false music to our ears. As an example, consider the so-called Mozart effect—namely, the proposition that being exposed to the soothing strains of Mozart and his classical pals in all venues of human waking, and even sleeping, activity, can have profound positive consequences on one’s intellectual skills and development. As a corollary of this musical interlude, a few years ago newspapers across the country reported the results of a research study that found that inner-city second graders who took piano lessons and received exercises that engaged their spatial ability experienced an improvement in their mathematics skills. Fascinating enough, but how much stake should the average consumer place in such a conclusion?

Everyone reading this article knows that the answer critically depends on the quality of the research conducted and the evidence obtained from it. Thus, how can we be confident that whatever improvements in math skills were observed resulted from students practicing the piano and computer-based spatial exercises rather than from something else? Indeed, the implied causal explanation is that such practice served to foster the development of certain cognitive and neurological structures in the students, which in turn improved their math skills: As one newspaper stated, “When children learn rhythm, they are learning ratios, fractions and proportions. … With the keyboard, students have a clear visual representation of auditory space” (Deseret News, 1999).

In the same newspaper account, however, other researchers offered alternative explanations for the purported improvement of musically and spatially trained students, including the enhanced self-esteem they may have experienced from such training and the positive expectancy effects communicated from teachers to students. Thus, at least in the newspaper version of the study, the evidence offered to support the preferred cause-and-effect argument is not compelling. Moreover, a review of the primary report of the research reveals that in addition to the potential complicators just mentioned, several methodological and statistical concerns seriously compromise the credibility of the study and its conclusions, including nonrandom assignment of either students or classrooms to the different intervention conditions, student attrition throughout the study’s 4-month duration, and an inappropriate implementation and analysis of the classroom-based intervention.

The possibility that music instruction combined with training in spatial reasoning improves students’ math skill is an intriguing one (see Chamberlin, 2003; but also Ho, Cheung, & Chan, 2003) and one to which I personally resonate. Yet, conclusions and recommendations emanating from the actual research as operationalized and conducted range from the sublime to the ridiculous. For example, colleague Diane Halpern (2000) describes the kind of “responsible reporting” that appeared in a popular science magazine for general consumers:

Preschool music activities such as singing “Twinkle, Twinkle Little Star” [might just be the] “brain food” that can help to develop future scientists and mathematicians because the pattern of neural firing involved in singing is similar to that used in mathematics and science. (p. 177)

To that I would add: “Oh, dear, what can the matter be?” Let me restate this in terms of the down-to-earth advice Elizabeth Loftus (1998), expert on eyewitness testimony and then-president of the American Psychological Society, offered to graduating seniors in her 1998 university commencement address:

There’s a wonderful cartoon that appeared recently in Parade magazine. … Picture this: mother and little son are sitting at the kitchen table. Apparently mom has just chided son for his excessive curiosity. The boy rises up and barks back, “Curiosity killed what cat? What was it curious about? What color was it? Did it have a name? How old was it?” I particularly like that last question. … Maybe the cat was very old, and died of old age, and curiosity had nothing to do with it at all. … My pick for the one advice morsel is simple: remember to ask the questions that good psychological scientists have learned to ask: “What’s the evidence?” and then, “What EXACTLY is the evidence?” (p. 27)

Stephen Stigler (1999) has written a fascinating treatise on the history of statistical methods and concepts, if it is permissible to use fascinating and statistical methods in the same sentence. In his book, Stigler relates a story about the renowned Karl Pearson, a forefather of the discipline of statistics. In London in 1910 arguments were being put forth concerning the deleterious consequences of parental alcoholism on their children’s development and subsequent behavior. But that’s exactly what they were—arguments—for as Stigler puts it, “… the argument that children of alcoholic families were the worse off for the alcoholism of their parents rested on no more than anecdotal evidence” (p. 15). The lack of evidence is what prompted Pearson to write to the London Times with a request.
of those making the argument to put the “statistics on the table, please” (p. 27), which would include a thorough description of the samples, methods, and measures used to produce the evidence. With no such evidence forthcoming, Pearson conducted his own empirical study of the problem, into which he was able to insert some of his own data-analytic contributions, including partial correlations that statistically accounted for various family demographics. For readers who care, Pearson’s study found little or no evidence to support the claim that parental drinking begets offspring problems. Pearson wanted to put the statistics on the table. When deciding whether claims about educational-intervention effectiveness are based on evidence that is credible, my favorite line has been, to paraphrase Cuba Gooding, Jr., “Show me the data!”

And whence exactly springs credible evidence? It comes from the conduct of credible research, which in turn follows directly from scientifically accepted methodological precepts. The essence of both scientific research and credible research methodology can in turn be reduced to the four components of what I have referred to as CAREful (Comparison, Again and again, Relationship, and Eliminate) intervention research and that colleague Sharon Derry and I drew from to provide a critical foundation for a first-year college course designed to improve prospective teachers’ scientific reasoning (see Derry, Levin, Osana, Jones, & Peterson, 2000). In particular, one can argue that evidence linking an instructional intervention to a specified educational outcome is scientifically convincing if: (a) the evidence is based on a Comparison that is appropriate (for example, comparing the intervention with an appropriate alternative or nonintervention condition); (b) the outcome is produced by the intervention Again and again (namely, that the outcome has been replicated, initially across participants in a single study and ultimately through independently conducted studies); (c) a direct Relationship (i.e., a connection or correspondence) exists between the intervention and the outcome; and (d) all other reasonable competing explanations for the outcome can be Eliminated (typically, through randomization and methodological care). Succinctly stated: If an appropriate comparison reveals again and again evidence of a direct relation between an intervention and a specified outcome while eliminating all other competing explanations for the outcome, then the research yields scientifically convincing evidence of the intervention’s effectiveness. In short, an ounce of credible evidence requires a pound of experimental control!

EDUCATIONAL-INTERVENTION RESEARCH, MEET MEDICAL RESEARCH!

An editorial in the New England Journal of Medicine clearly states the unacceptability of admitting anecdotes, personal testimony, and uncontrolled observations when evaluating the effectiveness of a new drug or medical treatment:

If, for example, the [New England] Journal of Medicine were to receive a paper describing a patient’s recovery from cancer of the pancreas after he had ingested a rhubarb diet, we would require documentation of the disease and its extent, we would ask about other, similar patients who did not recover after eating rhubarb, and we might suggest trying the diet on other patients. If the answers to these and other questions were satisfactory, we might publish a case report—not to announce a remedy, but only to suggest a hypothesis that should be tested in a proper clinical trial. In contrast, anecdotes about alternative remedies (usually published in books and magazines for the public) have no such documentation and are considered sufficient in themselves as support for therapeutic claims. Alternative medicine also distinguishes itself by an ideology that largely ignores biologic mechanisms, often disparages modern science, and relies on what are purported to be ancient practices and natural remedies. … Healing methods such as homeopathy and therapeutic touch are fervently promoted despite not only the lack of good clinical evidence of effectiveness, but [also] the presence of a rationale that violates fundamental scientific laws—surely a circumstance that requires more, rather than less, evidence. (Angell & Kassirer, 1998, p. 839)

The editors of the New England Journal of Medicine call for scientifically based evidence, not intuition, superstition, belief, or opinion. Many would argue that educational research is not medical research and that the former represents an inappropriate analog model for the latter. I disagree. Both medical and educational research involve interventions in complex systems in which mapping out causal relations is difficult. To illustrate, let us reread the first part of the New England Journal of Medicine excerpt, but this time from an educational-intervention perspective:

If, for example, the Journal of Educational Psychology were to receive a paper describing amelioration of a student’s reading disability after he had ingested a rhubarb diet, we would require documentation of the reading disability and its extent, we would ask about other, similar students whose reading did not improve after eating rhubarb, and we might suggest trying the diet on other students. If the answers to these and other questions were satisfactory, we might publish a case report—not to announce a remedy, but only to suggest a hypothesis that should be tested in a CAREful educational experiment. …

Just as medical research seeks prescriptions, so does educational-intervention research; and prescription seeking should be accompanied by scientifically credible evidence to support those prescriptions. As revealing countercases in point, let us turn briefly to the annual meetings of certain professional educational research associations. A session at one such annual meeting consisted solely of two research presenters displaying their wares in a joint presentation: Re-
Before continuing with the main thrust of my message, I sin-
guery, along with their resulting forms of "evidence," that are
glue out for critical examination two methods of empirical in-
search? Imagine the following dialogue: "Should the FDA ap-
prove the new experimental drug for national distribution?"
"Definitely! Its effectiveness has been documented in a poem
by one satisfied consumer and in a painting by another."

Educational historian Carl Kaestle (1993) makes no bones
about referring to the "awful reputation" of educational re-
search. Is it any wonder? Why is the medical research model
not good enough for educational research? As former Ameri-
can Educational Research Association president Michael
Scriven (1960) asks: "Is aspirin no longer working?" Econo-
mist Gary Burtless (2002) adds, "Why has randomization
been used to evaluate welfare policy, adult training, unem-
ployment insurance, job placement, medical interventions,
and a host of other policies … [but] controlled experimenta-
tion [is] almost never used to evaluate educational policy?"
(p. 179). Medical researcher Dr. Kenneth Anderson argues
that he sees a pressing need to move new drugs and therapies
from the "bench" to the "bedside." And so is there anything
wrong with creating an education counterpart for moving
new instructional treatments that are based on research evi-
dence that is "credible" to the "classroom"?

In a recently edited book, a National Research Council
Committee on Scientific Principles for Education Research
pushes the “good enough” envelope even further:

As our work began, we attempted to distinguish scientific in-
vestigations in education from those in the social, physical,
and life sciences by exploring the philosophy of science and
social science; the conduct of physical, life, and social sci-
ence investigations; and the conduct of scientific research on
education. We also asked a panel of senior government offi-
cials who fund and manage research in education and the so-
cial and behavioral sciences, and a panel of distinguished
scholars from psychometrics, linguistic anthropology, labor
economics and law, to distinguish principles of evidence
across fields. … Ultimately, we failed to convince ourselves
that at a fundamental level beyond the differences in special-
ized techniques and objects of inquiry across the individual
sciences, a meaningful distinction could be made among so-
cial, physical, and life science research and scientific re-
search in education. At times we thought we had an example
that would demonstrate the distinction, only to find our hy-
pothesis refuted by evidence that the distinction was not real.
(Shavelson & Towne, 2002, p. 51)

TWO CONTEMPORARY FORMS OF
EDUCATIONAL-INTERVENTION RESEARCH
INQUIRY

Before continuing with the main thrust of my message, I sin-
gle out for critical examination two methods of empirical in-
quiry, along with their resulting forms of “evidence,” that are
thriving in educational research today—even in educa-
tional–psychological research today. These are the
demonstration study and the design experiment.

The Demonstration Study

Two ubiquitous manifestations of demonstration studies in
educational contexts include: (a) an instructional interven-
tion that is introduced within a particular classroom with or
without a nonintervention comparison classroom; and (b) an
out-of-classroom special intervention program that is pro-
vided to a particular group of students. The critical issue here
(which will be revisited shortly) is that with only one class-
room receiving special instruction or only one group partici-
pating in a special program, separating the effects of the in-
tervention or the program from the specific implementation
of it is not possible.

Mary Levin and I discuss interpretive concerns associated
with the “evidence” derived from a demonstration study in
the context of evaluating the outcomes of an academic reten-
tion program (Levin & Levin, 1993). They are encompassed
in three CAREful-component questions in one: Was the pro-
gram effective? With an emphasis on effective, one can ask,
“Relative to what?” because in many program evaluation
studies frequently lacking is an appropriate comparison (ei-	her comparable nonprogram students or historical baseline
data). With an emphasis on the, one can ask: “Do you mean
this single implementation of the program?” because gener-
alization to other program cohorts or sites is not possible
without an again and again replication component. Finally,
with an emphasis on program, one can ask: “Can other,
nonprogram-related factors account for the observed out-
comes?” because without program randomization and con-
trol, one cannot readily eliminate other potential contributors
to the effects. A subsequent report of a college retention pro-
gram that we developed for academically at-risk minority
students (Levin, Levin, & Scalia, 1997) provides an example
of a demonstration study. Because of the uncontrolled nature
of the study and the one-time implementation of the program,
any of the documented positive outcomes associated with
program participants cannot be regarded as either scientifi-
cally credible or generalizable to other implementations
of the program. In that sense, then, and as we point out, a report
of that particular program and its outcomes can indicate only
what happened under a unique and favorable set of circum-
stances. It clearly is not an indication of what to expect if a
similar program were to be implemented by others with other
college students elsewhere.

The Design Experiment

Next let us consider a currently popular form of research in-
quiry in our field—through what has come to be known as the
design experiment. Alan Collins (1992) and the late Ann
Brown (1992) initially popularized the classroom-based de-
design experiment, which Gabi Salomon (1995), Bob Calfee (1992), and various research-funding agencies subsequently welcomed into the educational research community. In design experiments: (a) research is conducted in authentic contexts (e.g., in actual classrooms in collaboration with teachers and other school personnel), and (b) the experiment is not fixed in the traditional sense, but rather instructional-design modifications are made as desired or needed.

Interestingly and reminiscent of my earlier random sample discussion, I am inclined to regard design experiments as neither a design nor an experiment, although colleague Rich Lehrer has convinced me that they may be designed in the engineering sense of the word, rather than in the methodological sense. And they may also be beautifully designed in an aesthetic sense. In particular, in conventional research usage, design refers to a set of preexperimental plans concerning the specific conditions, methods, and materials to be incorporated in the study. In a design experiment, however, any components may be altered by the researcher or teacher as the investigation unfolds, as part of what Collins (1992) calls “flexible design revision.”

Similarly, in conventional research terminology, experiment refers to situations in which participants are randomly assigned to the two or more systematically manipulated and controlled conditions of a study. In a design experiment, however, appropriate randomization and control are conspicuously absent—which, in turn, do not permit a credible attribution of outcomes to the instructional procedure(s) under investigation.

The design experiment surely has its pros and cons. In contrast to the typical laboratory-based experiment, the design experiment is, by definition, classroom based and classroom targeted. On the other side of the ledger, design experiments can be criticized on methodological grounds. In my view, design experiments can play an informative role in preliminary stages of instructional development research—as long as the design experimenter remembers that the research was designed to be preliminary when reporting and speculating about a given study’s findings. My bottom-line plea here is for instructional researchers to think at least as much about designing experiments as they think about design experiments.

ALL THE RESEARCH WORLD’S A STAGE

With this backdrop, demonstration studies, design experiments, and other informal classroom-based investigations are incorporated into the model of educational-intervention research that Levin and O’Donnell (1999) proposed. The model incorporates as a critical classroom-based instructional-research stage the systematic conduct of what we have imported from the medical-research model as randomized classroom trials. Levin and O’Donnell ask: “What [should we] do about educational research’s credibility gaps?” Our vision of how to close a fundamental credibility gap while better informing instructional practice is presented in the following stage model of educational-intervention research.

Stages 1 and 2 of the model are probably very familiar to educational–psychological researchers because studies in those traditions comprise the vast majority of existing educational research. Stage 1 consists of preliminary ideas, hypotheses, observations, and pilot work, whereas Stage 2 is represented by controlled laboratory experiments, on the one hand, and classroom-based demonstrations and design experiments, on the other. Both controlled laboratory experiments and classroom-based studies are preliminary, although in different complementary senses. The former are preliminary in that their careful scrutiny of instructional interventions lacks a classroom-implementation component, whereas the latter are preliminary in that their intervention prescriptions are often not founded on scientifically credible evidence. Stage 2 research also encompasses observational studies of master teachers deploying their craft knowledge in longer term teacher–researcher collaborative investigations, as envisioned by Hiebert, Gallimore, and Stigler (2002). Stages 1 and 2 studies are crucial to developing an understanding of the phenomena that inform classroom practice but that first must be rigorously, complexly, and intelligently evaluated in Stage 3 (randomized classroom trials studies). Failure to consider possibilities beyond Stages 1 and 2 may result in a purposelessness to research, a temptation never to go beyond understanding a phenomenon and determining whether it is a reliable phenomenon with genuine instructional implications. The accumulation of classroom-based scientifically credible evidence is precisely the function of the randomized classroom trials stage of the model. As in medical research, this consists of an examination of the proposed treatment or intervention under realistic, yet carefully controlled, conditions.

Realistic conditions refer to the specific populations and contexts about which one wishes to offer conclusions regarding treatment efficacy, namely, external validity desiderata. In medical research, the conditions of interest generally include humans rather than animals, whereas in educational research, the conditions of interest generally include students in classrooms rather than isolated individuals. In addition, in both medical and educational contexts, the interventions (e.g., drugs or instructional methods, respectively) must be administered in the appropriate fashion (e.g., dosage levels or instructional integrity, respectively) for a long enough duration for them to take hold and to permit the assessment of both the desired outcome (e.g., an improved physical or academic condition, respectively) and any unwanted side effects (e.g., adverse physical, cognitive, affective, or behavioral consequences). In a classroom situation, an appropriately implemented instructional intervention of at least a semester, or even a year, in duration would be expected to satisfy the “long enough” criterion.
Carefully controlled conditions refer to internally valid experiments based on the random assignment of multiple independent units to alternative treatment or intervention conditions. Again, in medical research, the randomized independent units are typically people, whereas in educational-intervention research, the randomized independent units typically ought to be classrooms or schools. As with medical research, careful control additionally involves design safeguards to help rule out contributors to the effects other than the targeted intervention, such as including appropriate alternative interventions, incorporating “blind” and “double blind” intervention implementations—to the extent possible—to eliminate student, teacher, and researcher biases, and being responsive to all other potential sources of internal invalidity, such as selective (nonrandom) classroom or student-within-classroom dropout. How’s that as a strong hint to your earlier homework assignment?

The randomized classroom trials stage of this model is sensitive to each of the earlier discussed CAREful research components in that: (a) the inclusion of alternative interventions permits meaningful comparison when assessing the effects of the designated intervention—which represents a significant consideration in light of recent media coverage documenting the power of placebos; (b) the use of multiple independent classrooms, both within a single study and, ideally, as subsequent replication studies, permits generalization through the specified outcomes being produced again and again; (c) with across-classroom randomization of interventions, and assuming adequate control and appropriate implementation of them, whatever relation is found between the targeted intervention and the specified outcomes can be traced directly to the intervention; because (d) with such randomization, control, and implementation, one is better able to eliminate all other potential explanations for the outcomes.

In educational research, one of the most pervasive impediments to securing scientifically credible evidence relates to the manner in which classroom-based research is typically conducted and analyzed. Without going into details here, the critical issue concerns the need for what in the medical and prevention research literature is called cluster randomization and analysis when the associated research conclusions and inferences are applied to group-based interventions (see, e.g., the Web site of medical research’s prestigious Cochrane Collaboration at www.cochrane-net.org/openlearning/HTML/modA2–4.htm). A similar major scientific credibility problem pertains to researchers ignoring teacher and classroom effects (e.g., Levin, 1992a), which has consequences akin to those Wampold (2001) articulated in his discussion of therapist effects—as more plausible alternative explainers of therapy effects—in accounting for outcomes of psychotherapy (see also Serlin et al., 2003).

The randomized classroom trials stage of Levin and O’Donnell’s (1999) proposed model possesses the best of what carefully controlled and well-executed laboratory-based research has to offer classroom-based research. Foremost among these is the inclusion of multiple classrooms that are randomly assigned to receive either the designated instructional intervention or an acceptable alternative. The use of multiple independent classrooms is imperative for combating “evidence credibility” concerns arising from both methodological and statistical features of the research. Each of these features is discussed extensively in Levin and O’Donnell’s article (see also Levin, in press). My sentiments are also reflected in a 2002 book, Evidence Matters: Randomized Trials in Education Research, edited by influential statistician Frederick Mosteller and Robert Boruch (2002), an educational methodologist who has been promoting the cause for 30 years and who has documented hundreds of examples of randomized field experiments in the social sciences. The editors write:

Some of the best evidence to address the question [of which interventions work better] can be generated in randomized field trials … situations in which individuals or entire organizations are randomly assigned to one of two or more interventions. … [Accordingly,] the groups are statistically equivalent at the outset [and] we can then be assured that comparison of the relative effectiveness of the intervention will be fair. That is, the properly executed randomized field trial will produce estimates of relative effects that are statistically unbiased. Furthermore, one can make legitimate statistical statements about one’s confidence in the results. (p. 2)

Conducting randomized classroom trials studies is not an easy task. Nonetheless, randomized classroom experiments are not impossible, or even impractical, to conduct. Neither are the commonly offered objections to them difficult to counter, which Cook and Payne (2002) have effectively demonstrated recently with nine such contentions. Consequently, educational-intervention researchers must begin adding these to their investigative repertoires to enhance the scientific credibility of their research and research-based conclusions. There can be no denying that in contrast to the independent and dependent variables of the prototypical laboratory experiment, the factors related to school or classroom outcomes are complex and multidimensional. Yet, others have argued compellingly that to understand the variables and variable systems that have implications for social policy, randomized experiments should—and can—be conducted in realistic field settings. I am presenting a similar argument for more carefully controlled classroom-based research on instructional interventions and on other educational prescriptions.

Beyond the level of client- or student-directed interventions are classroom-, school-, and community-directed interventions that are supported by scientifically credible evidence. A superb accounting of nine such large-scale inter-

---

1Incredibly (or, in the current context, should I say “credibly”?), the multiple-classroom criterion is reflected in the emphasis-added portion of Thorndike’s (1910) reflections on ideally conducted educational-intervention research, which I quoted at the beginning of this article.
ventions may be found in Crane’s (1998) Social Programs That Work. Several of the 13 criteria Crane developed to assess whether a particular social program should be regarded as effective feature scientifically valid methods that yield convincing quantitative evidence—namely, outcomes that are both statistically nonchance and of a magnitude that can be considered practically worthwhile.

As was just noted, the social science literature is replete with examples of CAREfully conducted field experiments. In an educational research context, these have encompassed—and should continue to encompass—classroom- and school-trials experiments, including single-case designs based on various behavior-analysis and time-series models (e.g., Kratochwill & Levin, 1992). Examples of exemplary randomized classroom-, school-, and community-trials intervention research can be found in the Tennessee class-size study, the Hutchinson smoking-prevention project, the Fast Track conduct problems prevention study, and certain of the Success for All evaluations, to name just a few.

A BRIEF TIMEOUT FOR A FEW IMPORTANT DISCLAIMERS

And now a brief “disclaimers” timeout. First, I do not mean to give the impression that I advocate the universal adoption of randomized classroom-trials studies in all educational research and by all educational researchers. Not every educational research investigation should be of that scope or magnitude. Rather, only when an instructional method or variable has advanced to the level of consideration for public consumption should it adhere to the scientific credibility standards associated with randomized classroom-trials studies. Neither should every researcher be advised to conduct randomized classroom-trials studies. I certainly would not recommend that task, for example, for an assistant professor developing an individual research program on the road to academic tenure. In contrast, senior, secure researchers might be encouraged to undertake more ambitious classroom-trials investigations.

Second, allow me to throw in some food for additional thought in the form of a provocative methodological monkey wrench. Counseling psychology researcher, methodologist, and colleague Bruce Wampold (2001) has recently proffered a bold assertion that the so-called medical model represents a poor aspirational target for at least one domain to which an enormous amount of psychological research has been devoted: psychotherapy and, in particular, the search for effective treatments or therapies in that domain. Wampold’s contention is that the quest for unique therapies and therapy components has proven futile because specific therapy-related differences pale in comparison to a host of identifiable general psychotherapy factors in what he calls a contextual model. Such general or common factors include the patient’s belief in the efficacy of the therapy (akin to placebo effects), the therapist’s allegiance to the particular therapy being implemented, and the therapist–patient alliance, among others. In particular, Wampold provocatively argues on the basis of the massive empirical data he has examined that all bona fide psychotherapies are equally beneficial, given that the other positive common factors are equally operational (see also Klein et al., 2003).

But what do psychotherapy treatments have to do with educational–psychological interventions you may ask? Plenty, to the extent that in whatever educational–psychological domains the interventions under consideration can be shown to incorporate common factors, rather than identifiable unique components or active ingredients. That is, the medical model that I am advocating here is apropos only when one’s purpose is to investigate the efficacy of an educational-intervention, relative to either a no-intervention control or an alternative intervention in cases where it can be assumed or it has been established that the intervention under consideration consists of more than just general student receptiveness, motivational, attentional, and expectancy factors along with factors that are embedded in almost any substantively appealing intervention approach. The intervention implementer’s allegiance to, competence with, and enthusiasm for the particular approach would also fall into this common or intervention-unrelated factors category. Educational interventions targeted primarily at students’ academic motivation and behaviors, or their general sense of well being, might well fall into this class. Thus, the current discussion necessarily focuses on educational–psychological interventions that can be shown to include specific, theoretically based, or empirically based ingredients—for example, skill-demanding instructional methods or educational-intervention programs with agreed-upon specific components and for which common elements can be eliminated as the desired outcome producers.

THE “EVIDENCE-BASED” MOVEMENT: THREE CHEERS FOR CREDIBLE EDUCATIONAL-INTERVENTION RESEARCH!

The spirit of scientifically credible research is part of what has come to be known as the evidence-based movement in many research disciplines. The movement is well represented in the evidence-based guidelines for effective interventions that have been formulated by several American Psychological Association division task forces, as well as various foundations and collaborations (e.g., Campbell Collaboration, www.campbellcollaboration.org; What Works Clearinghouse, www.w-w-c.org); reports (e.g., Boruch et al., 2002); summer institutes; international conferences; and of serious pragmatic concern to educational psychology researchers, federal funding agencies. As one such example of the movement to incorporate the term scientific into educational research, take a look at some of the funded research projects and requests for proposals associated with the
still-in-its-infancy Institute of Education Sciences (e.g., Reyna, in press; Whitehurst, 2003). Quite a shocking—but for me, exciting—departure from the direction in which the American Educational Research Association has been headed the last couple of decades.

Evidence-based interventions are more likely to bear directly on a related problem as well: how to change the public’s perceptions of and confidence in educational research. Colleague Angela O’Donnell (2001) has commented on the public’s common perceptions of the everyday expertise and knowledge in the field of education—the even-my-grandmother-could-have-told-you-that syndrome—compared to those in more technical and scientific fields. To paraphrase Aretha Franklin, educational experts (including both educational researchers and educational practitioners) could benefit from a heavy dose of genuine R-E-S-P-E-C-T. Creating a knowledge in the field of education—the even-my-grandmother’s perceptions of and confidence in educational research.

In charting a research course, one sometimes finds (as poet Robert Frost once did) that “two roads diverge in a yellow wood” and that a choice must be made regarding the road on which to travel. Of course, personal preferences and rational thinking enter into those choices. But so do the less controllable factors of timing and luck—yup, here’s that ubiquitous randomness again! The critical role fortuitousness, or serendipity, plays in research is wonderfully illustrated in a classic case that I have my students read: “The case of the floppy-eared rabbits.” The case, which sociologists Bernard Barber and Renee Fox (1958) detailed, involves the accidentally discovered consistent drooping of laboratory rabbits’ ears following the injection of a certain laboratory substance—a phenomenon that one scientist noticed and pursued and another scientist noticed but did not pursue. In the apt words of Barber and Fox, serendipity found for one scientist is serendipity lost for the other.

Many of us can surely think back to research choice-points we made along the way, some of which may have panned out while others were just panned. As my own personal case of possible serendipity lost, years before reciprocal teaching, think-aloud protocols, and instructor modeling-and-feedback notions were in vogue, they were on my own laboratory drawing board—but were never developed to the point of systematic research exploration. Yet what fizzled on the laboratory drawing board was surely for me a case of serendipity found in that it enabled me to pursue other substantive and methodological and statistical topics for which I had enormous passion, which brings me to my second slice of advice to aspiring educational researchers.

Slice of Advice No. 2: Opt for Passion Over Fashion

Do the research about which you are most passionate, even if it is not trendy at the moment or on today’s federal-funding bandwagon. Sure, getting research funding for a trendy topic that you are passionate about is always great, but if you must opt for one, I would urge you to opt for passion over fashion! In company with serendipity, Alice Dreger (1999) colorfully asserts in a New York Times essay that “… passion can allow a scientist to surge past the clouds formed by accepted wisdom and reach new discoveries tamer colleagues would not.” She continues, in even more colorful language, that she knows “a senior physicist who is so passionate about quarks that he probably has erotic dreams about atoms, but no one doubts his ability as a physicist. On the contrary, his passion wins him awards and financing, and it probably should.”

My own more modest “passion” example comes from the substantive area that I have devoted the bulk of my research attention to over the past almost 40 years, applications of visual imagery, pictures, and memory-enhancing techniques to improve learning and instruction (see also Levin, 1996). My specialized interest in combining memory and systematic methods from an experimental research perspective comes from a fascination in understanding the conditions under which memory-enhancing methods “work” so that they can be transported in various forms to both students and other citizens who would greatly benefit from their use. My personal research passion has been directed at uncovering the different varieties of pictorial aids and devices that can be injected into this instructional enhancement process, while assessing the associated limits and limitations of such techniques.

Slice of Advice No. 3: And if at First You Don’t Succeed …

Perseverance in academic research and writing is a fundamental precept. Perseverance here refers to sticking with a research topic, project, or product in which you have confidence, and it includes aspects of both inquisitiveness (typically on the front end) and compulsiveness (typically on the back end) of a research-and-writing endeavor. In short, do not let setbacks discourage you. As an aside, although the
And a Bonus Slice of Advice

Let me throw in a little bonus slice of advice for you: Occasionally, explicitly acknowledging the others who contributed in various ways to your own research and teaching accomplishments is a good idea. I have tried to be diligent in doing that throughout my career; in fact, on occasion I have even been accused of being too diligent! Yet, acknowledging the input of others is good, and so I pass along similar encouragement to you: Be generous in your giving appropriate credit to the colleagues, students, and other players who are instrumental in helping you achieve your own academic successes. Exhibiting a little more humility than hubris will serve you well in this profession.

Finally, Thankfully!

Earlier I spoke of research passion. As important as passion is (particularly early in one’s career), it is not the sole purpose for pursuing a research topic. Another purpose is to provide hope and inspiration for seasoned educational researchers, myself included.

Although I disagreed with Berliner’s (1992) interpretations concerning the enduring contributions to our discipline of Thorndike and Dewey, I have no disagreement whatsoever with another argument that Berliner makes in the same paragraph of his 1992 article:

The stories we need to tell to exert more influence, must be entwined with the “stuff” of the sites in which we hope our data and concepts will prove useful. It is by following up our laboratory findings in real-world settings, or by using real-world analogues to illustrate our laboratory findings, that we encourage understanding of what we find, perhaps allowing us to influence the policymakers and teachers in the ways that we had hoped. By testing our ideas in real-world settings we gain the direct experience needed to promote those ideas. (pp. 157–158)

Three critical considerations in prescribing specific interventions (summarized in Table 1) capture what I consider the ideal to which educational researchers should aspire. Two of these I have referred to (Levin, 1994) as scientific credibility, which incorporates internal validity and statistical conclusion validity criteria, and educational creditability, which incorporates standards of societal importance (e.g., Wolf’s [1978] social validity). Whereas I formerly included notions of construct validity and external validity in the creditability category, I now place these in their own alliterative middle category of contextual accretability, or research accumulation and generalization efforts under varying conditions involving both procedural variations and student and situational differences. Notably, research that is both scientifically credible and contextually acceptable may or may not be educationally creditable. For me, “good” educational-intervention research embodies all three: It is credible, acceptable, and creditable.

### TABLE 1

<table>
<thead>
<tr>
<th>Research Criterion</th>
<th>Primary Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific credibility (internal validity, statistical conclusion validity)</td>
<td>Provides confidence that a study’s outcomes are truly produced by the intervention</td>
</tr>
<tr>
<td>Contextual accretability (construct validity, external validity)</td>
<td>Establishes the intervention’s scope of applicability</td>
</tr>
<tr>
<td>Educational creditability (social validity)</td>
<td>Assesses the potential value of the intervention to society</td>
</tr>
</tbody>
</table>
Several years ago at an informal social gathering, American Psychological Association Division 15 past-president Gabi Salomon—in his characteristically lively conversational style—challenged educational—psychology researchers to do more research and writing, not on educational topics with which we are familiar, expert, experienced, and feel comfortable, however, but on topics that have the potential to make a genuine difference to the lives of real teachers and real students with real curricular materials and real academic problems in real classrooms. Adopting my educational credibility construct and echoing the challenge, I urge my colleagues to accept the challenge and, to paraphrase Robin Williams: “Carpe millennium!” As for me, I aim to continue to do whatever is humanly possible in the way of proselytization on behalf of enhancing the scientific credibility of educational research.

ACKNOWLEDGMENTS

This article is based on an American Psychological Association Division 15 Thorndike Award address delivered at the American Psychological Association annual meeting in August 2002.

I am eternally grateful to the numerous colleagues and students who worked so hard to help further my research career over the years.

REFERENCES


Newcombe, N. S. (2002). Biology is to medicine as psychology is to education: True or false? In D. F. Halpern & M. D. Hakel (Eds.), Applying the science of learning to university teaching and beyond (pp. 9–18). San Francisco: Jossey-Bass.


