Articles should deal with topics applicable to the broad field of program evaluation. Articles may focus on evaluation methods, theory, practice, or findings. In all cases, implications for practicing evaluators should be clearly identified. Examples of contributions include, but are not limited to, reviews of new developments in evaluation, descriptions of a current evaluation study, critical reviews of some area of evaluation practice, and presentations of important new techniques. Manuscripts should follow APA format for references and style. Length per se is not a criterion in evaluating submissions.

Grades of Evidence

Variability in Quality of Findings in Effectiveness Studies of Complex Field Interventions

Madhabi Chatterji
Teachers College, Columbia University

Abstract: This article argues with a literature review that a simplistic distinction between strong and weak evidence hinged on the use of randomized controlled trials (RCTs), the federal "gold standard" for generating rigorous evidence on social programs and policies, is not tenable with evaluative studies of complex, field interventions such as those found in education. It introduces instead the concept of grades of evidence, illustrating how the choice of research designs coupled with the rigor with which they can be executed under field conditions, affects evidence quality progressively. It argues that evidence from effectiveness research should be graded on different design dimensions, accounting for conceptualization and execution aspects of a study. Well-implemented, phased designs using multiple research methods carry the highest potential to yield the best grade of evidence on effects of complex, field interventions.

Keywords: *evidence standards; evidence-based practices; mixed-method designs; randomized experiments*

Purpose

The current federal policy drive for evidence-based models of practice in wide-ranging public service sectors has been coupled with a concurrent push for the use of randomized controlled trials (RCTs) as the "gold standard" for generating rigorous scientific evidence on whether or not field-based practices, programs, policies, and interventions work (referred to

Madhabi Chatterji, PhD, Associate Professor of Measurement, Evaluation and Education, Codirector, Assessment and Evaluation Research Initiative (AERI), Teachers College, Columbia University, Box 68, 525 West 120th Street, New York, NY 10027; e-mail: mb1434@columbia.edu

American Journal of Evaluation, Vol. 28 No. 3, September 2007 239-255
DOI: 10.1177/1098214007304884

as "programs" hereafter). In education, the federal stance on the standard for science was articulated through influential legislation such as the No Child Left Behind Act of 2001 (NCLB; 2002), the Education Sciences Reform Act of 2002 (ESRA), and related government publications.

The basic principle that social programs and policies are supported by proven research is unquestioned. However, the declared preference for one type of research design (namely, RCTs) over others has unsettled researchers and evaluators in fields as diverse as education, psychology, and health. Serious discussions continue in education as the challenges in implementing laboratory-style RCTs under field conditions are documented (see, e.g., Berliner, 2002; Feuer, Towne, & Shavelson, 2002; Levin & O'Donnell, 1999; Schoenfield, 2006; Wolff, 2000).

Seeking out better methods for the conduct of research and evaluation in field settings is but one of many necessary challenges in the present policy climate. To promote the adoption of evidence-based models on a large scale, there is a concomitant need to educate potential users as to what constitutes research evidence of acceptable quality. For some time now, consumers of effectiveness research—other researchers, evaluators, funders, policy makers, and practitioners—have received a singular message from high-level policy makers concerned with evidence-based models of practice: namely, that RCTs generate "strong evidence" of what works; all other designs yield meaningless or largely untrustworthy data on impact questions (see, e.g., U.S. Department of Education, 2003, p. v). How well does the cited distinction between "strong" and "weak" evidence tied to RCT use serve the cause for evidence-based practices when "socially complex" (Wolff, 2000) service interventions are the object of study?

In this article, I posit that a simplistic distinction between strong and weak evidence on the sole basis of RCT use is not a tenable proposition when complex social programs are the focus of evaluative inquiry. Using a multidisciplinary literature review, I demonstrate that in such cases the quality of evidence from singular reliance on RCTs suffers because of researchers' inability to fully explain outcomes or rule out alternative explanations of effects due to a lack of situated knowledge relevant to the developmental stage of a program and its environment. Even in cases where RCTs are scaffolded with some level of post hoc[1] data gathering on implementation and contextual variables, the dearth of a comprehensive body of evidence collected over the life of a program undermines the meaningfulness and utility of the evidence on effects. Instead, the likelihood for obtaining a better caliber of evidence on the effects of complex interventions is improved greatly by employing *phased*, *mixed-method designs* that incorporate delayed experiments. Such alternative designs—labeled recently as *extended term, mixed method* (ETMM) type of designs (Chatterji, 2004)—are guided by a broader set of research questions delving into not simply whether or not a program has an effect that is better than chance but also the nature, breadth, and sustainability of the effects observed. The currently touted notion of a hierarchy of research designs associated with the quality of research evidence, with RCTs placed right at the top, thus needs serious reconsideration.

Specifically, I introduce the concept of *grades of evidence*, likened to grades of refined oil products, adding criteria over and above the traditionally used criteria of effect size and statistical significance of findings (as given in federal policy tools such as the What Works Clearinghouse Study DIAD[2]), for evaluating evidence quality. I argue that effectiveness studies of complex social programs and policies, such as those typically found in education, should be classified with regard to the "grade of evidence," using different and expanded criteria for evidence appraisal, as even RCTs vary with regard to evidence quality. The quality of evidence varies depending on field conditions, complementary methodological choices made by researchers, and quality of execution. This article urges the research and evaluation

community to move toward thinking about "grades of evidence" as opposed to a dichotomy of "good" versus "bad" evidence hinged to RCT use.

Utility of RCTs When Studying Complex Programs and Policies: What Constitutes "Strong Evidence" of Impact?

Some Definitions and Starting Premises

Social programs, referred to as *field-based programs* herein, comprise multiple, human, social, and operational elements and operate in some larger organizational or community setting. An effectiveness study, whether conducted under the rubric of academic or evaluative inquiry,[3] seeks to answer questions on the impact of field-based programs, policies, and interventions on some targeted outcome(s). When conducted as evaluation research, such studies simultaneously aim to inform stakeholder needs and aspire more directly toward improving social conditions and policies.

The quality of scientific evidence generated through effectiveness studies on field-based programs is contingent on appropriate choice of research design(s) vis-à-vis the questions asked. Furthermore, it also depends on how well the chosen design(s) can be and are executed under field conditions, given the constraints faced by researcher(s). In any study, methodological actions and decisions of researchers and field conditions under which they themselves work influence the quality of evidence it yields.

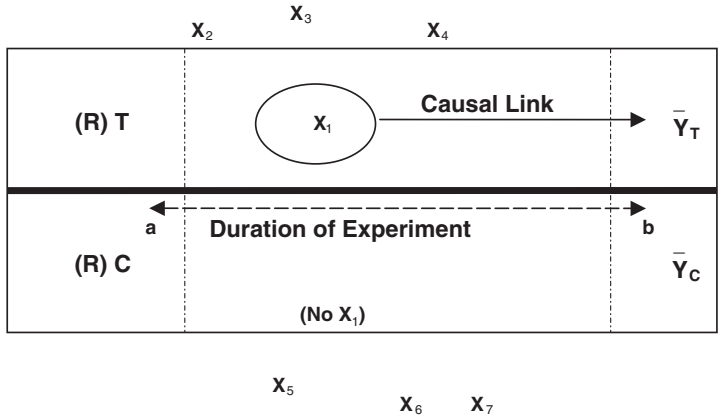
I begin with the premise that, when studying field-based programs, it is difficult if not impossible for impact studies to generate evidence on net effects on targeted outcomes. Net effects are effects of an intervention on desired outcomes with effects of other variables held constant, or otherwise controlled and accounted for. Gross effects are total effects inclusive of effects of other confounders or extraneous factors other than the intervention under study.

However, it is possible for researchers to aspire toward the highest "grade of evidence" with deliberate design actions and choices. This section will substantiate the argument as to why RCTs used alone fail researchers in their cause to obtain the highest grade of evidence on the effects of field-based programs; why quasi-experiments by themselves are less useful, as they are initiated with design handicaps that can rarely be overcome or compensated for once an empirical study is under way; and why complementary use of other research methods with comparative experiments, attending to various intervention components, the alternate conditions, and systemic factors over the developmental life of a program, can enhance the overall grade of evidence on program effectiveness.

Merits of RCTs

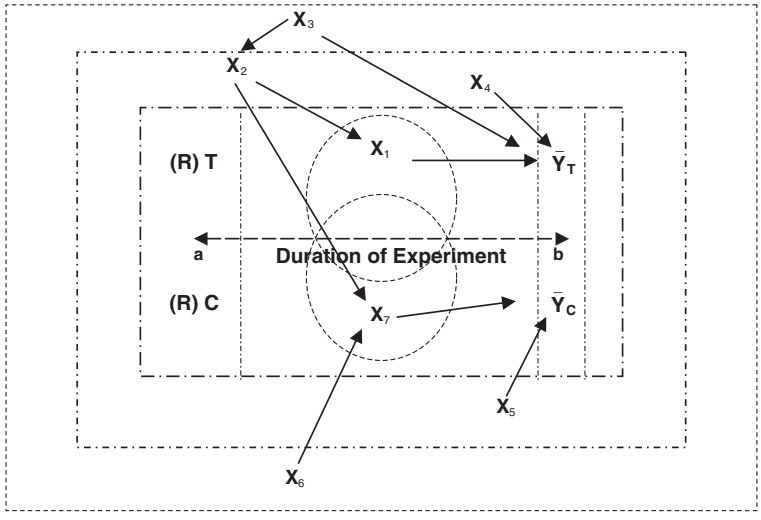
RCTs are appropriate designs when the aim is to ascertain causal effects, or to establish a causative link between a manipulated intervention, *X*, on some desired, observable outcome, *Y*, in a given population, all else being equal. Starting with a defined population, the basic RCT design is implemented by initially setting up two equivalent groups through random assignment of individuals (cases) to the a new program (Treatment, *T*), and an alternate condition (Control, *C*). Through randomization, RCTs provide researchers with a definitive strategy to equalize preexisting differences in T and C groups on variables that could eventually affect the outcome, capitalizing on principles of probability and independence of case assignment. When all relevant extraneous or confounding variables can be equalized through randomization, individuals in both groups are expected to display the

Figure 1
Theoretical Assumptions of Randomized Controlled Trials (RCTs)



Legend
X_1 = Intervention/under study
$\bar{Y}_C, \bar{Y}_T$ = Means on desired outcome in T and C groups
$X_2 - X_7$ = Extraneous variables
T, C = Treatment and control groups
(R) = Randomly assigned subjects to T and C groups

Figure 2
Typical Field Conditions for Randomized Controlled Trials (RCTs)



Legend
X_1 = Intervention/under study
$\bar{Y}_C, \bar{Y}_T$ = Means on desired outcome in T and C groups
$X_2 - X_6$ = Extraneous variables; X_7 = Alternate program in control setting.
T, C = Treatment and control groups
(R) = Randomly assigned subjects to T and C groups

same average outcome if their group assignment were switched. In sum, if RCT designs meet theoretical expectations, conditions are optimized in the T and C groups for making causal connections between X and Y. Meeting the two assumptions of randomness and independence also creates optimal conditions for making statistical inferences using significance tests, which are also governed by the laws of chance.

The question remains: Do RCTs on field-based programs generate the best "grade of evidence" on the causal link between X and Y by virtue of the random assignment procedure (assuming it is well executed), as claimed in federal documents and elsewhere? To demonstrate the numerous threats to an idealized, laboratory-type RCT under field conditions, consider a typical evaluation case in education (see Figures 1 and 2).

The Ideal Versus the Actual: Field Experiments in Education

The scenario concerns an education program targeting an elementary school population. Let us suppose researchers are investigating the effects a treatment variable—an innovative reading program, X_1—on reading achievement, Y (the desired outcome in this case), typically assessed at the end of each school year. To start, researchers gather adequate information about the program and the schools where it is being tested.

Let us say the program has been added on to strengthen an existing, basal reading curriculum. Contextually, the students are from schools in a large district that is trying hard to comply with current No Child Left Behind Act of 2001 requirements (2002), such as recruitment of qualified teachers, providing additional professional development opportunities to teachers, offering supplementary instruction to struggling students after school hours, and providing English as a Second Language (ESL) resource teachers in classrooms serving nonnative speakers of English.

The ideal field experiment. Figure 1 offers a diagrammatic view of what a theoretically sound RCT design would look like, if implemented in the context described, to examine effects of the reading innovation, X_1. To start, the new program would be randomly assigned to an adequate number of students, satisfactorily equalizing background factors that might differentially influence the outcome *Y*, and meeting the criteria of independence and chance in how participants are distributed in T and C groups. The number of students would be large enough in both conditions to create optimal conditions for statistical efficiency and power, with an analytic plan accounting for nesting of students in classrooms that may violate the assumption of independence. The planned experiment would span a sufficiently long period for the treatment to have potency when delivered per plan, say, a 9-month school year or a semester, extending from Point a to Point b in Figure 1. Once assigned, the T and C groups would be clearly separated in solidly bounded regions, shown as upper and lower rectangles in Figure 1, where the numbers of participants in each group would be held intact throughout the duration of the experiment. There would be a tight, closed boundary around the groups, indicated by the two rectangles, keeping extraneous variables and confounders, denoted as X_2, X_3, X_4 . . . X_7, in Figure 1, out of the experiment as long as it continues. Furthermore, X_1, the treatment with all its components, would function as a coherent, defined whole, just as expected by developers—or, in evaluation terms, the program would be implemented per the logical plan of action or underlying program theory. The treatment, X_1, would also be clearly bounded and effectively manipulated (denoted with a tight, solid circle in bold). *Effective manipulation* indicates that the C children would not receive any part of X_1, the new reading program, at all—there would be no leakage or contamination. What occurs in reading in the upper rectangle would remain operationally distinct from that in the lower one. Finally, the

experiment would continue under this set of conditions for the period marked (a–b), allowing for outcomes to be measured and averages compared between the two groups in a scientifically meaningful manner at its conclusion. If significant changes are documented in the desired direction in T children following the experiment, a causal inference linking the new program to the outcome would now be conclusive.

An actual field experiment. Now consider Figure 2, depicting how actual field conditions in a typical school violate assumptions for implementing the laboratory-style RCT depicted in Figure 1. The rectangular boundaries around the T and C groups are now dotted and broken, not solid, as most school organizations are open systems. Children originally assigned to experimental groups may move in and out of the marked zones during the course of the experiment because of family mobility, administrative actions, or other factors unrelated to the experiment. Broken boundaries will upset the randomization process and typically result in differential attrition rates from T and C conditions. Although students are in their delineated "experimental" and "control" spaces, they are nested in a hierarchical organizational structure (e.g., students nested in classrooms, which are nested in schools, which belong in larger districts or regional communities). Furthermore, the broken boundaries around the T and C spaces allow many outside factors to seep in and affect outcomes in different ways.

Outside children's background characteristics that were intentionally randomized at the starting point, as shown, many differentially distributed outside factors may have direct, mediating, moderating effects on T and C groups' outcomes. For example, X_2 (say, a school-wide emphasis on literacy), X_3 (district-wide teacher training programs on standards-based reading instruction), X_4 (need-based ESL supports in classrooms), or X_5 (optional child participation in after-school supplemental reading programs) are a few variables that can mediate or moderate student performance on the same outcome measure. Because of layering of the new program on an existing one that is continuing in T and C conditions, the treatment is not distinct from the control condition at all. Rather, it is overlapping and loosely defined in the early stages of program implementation (shown with overlapping dotted circles). Furthermore, it is confounded with teacher quality and available instructional resources that T and C teachers can opt to use as they deem fit.

Despite the randomized experiment continuing for the specified period in Figure 2, effects found on its conclusion are particularly difficult to interpret if only outcome differences are compared and evaluated by group. Without adequate and supporting data on context variables, implementation inputs, delivery processes, and other factors carrying the potential to affect outcomes simultaneously, outcome results are difficult to link with the new intervention, as well as hard to understand, explain, and defend.

Implications of Violated RCT Assumptions

What are the fallacies in an overreliance on RCT designs under conditions depicted in Figure 2, even if we assume that randomization is well executed and outcomes defensibly selected and measured? Although the preceding case is billed a hypothetical one, field researchers will instantly recognize the scenario as quite typical in education. Assuming that randomization was properly done, we find that the major threats and interferences come into play during the course of the field experiment, or *after* random assignment is completed, and data-gathering and analytic decisions are made. If we further assume high levels of implementation fidelity and effectiveness in manipulation of the treatment, we find that a lack of formal information on environmental factors likely to interact with or mediate the treatment and influence the outcome (and which cannot realistically be controlled in field settings) still

undermines the quality of the evidence on program impact. Effects are typically "gross" rather than "net"—but because traditional RCTs generate little or no information on factors intervening with or confounding the manifested effect, emphasizing instead group differences on mean outcomes, the true nature of the effect remains masked.

Design Choices and Actions to Improve the Grade of Evidence From Experiments

Can some of the above threats be addressed by conscious design actions, thereby improving the grade of evidence? The major RCT assumptions of group equivalence and independence call for empirical checks and appraisal once a field experiment is in progress. The first assumption of equivalent distributions on relevant background variables holds true only if the size of the samples in treatment and control conditions is large enough vis-à-vis the number of characteristics that need to be evenly distributed. Once that distributive balance is achieved per some a priori but defensible criterion set by researchers, the sample must remain unchanged over the course of an experiment in terms of size and composition. One cannot simply assume that groups remain equalized because, empirically, the assumption is rarely realized. Empirical verification is thus necessary, not only at the start of an experiment but also at strategic points as a study continues so that violations can be detected and evaluated from a descriptive standpoint. This practice is rare or uneven, as may be apparent through a cursory review of published experiments across education, health, and psychology journals.

Periodic descriptive checking of the equivalency of T and C samples on relevant variables throughout an experiment was the recommended practice of the pioneering "father" of the true experiment. In the classic, *The Design of Experiments*, R. A. Fisher (1960) acknowledged the preceding challenges in equalizing treatment and control conditions during randomization, urging researchers to recognize that

> whatever the degree of care and experimental skill . . . expended in equalizing the conditions, other than the one under test, . . . this equalization must always be to a greater or lesser extent incomplete, and in many important practical cases will certainly be grossly defective. We are concerned, therefore, that this inequality whether it be great or small, shall not impugn the exactitude of the frequency distribution, on the basis of which the result of the experiment is to be appraised. (p. 19)

Fisher thus asked us to go beyond statistical significance (based on rejection of a null hypothesis) and effect size to interpret the results of a randomized experiment. To better evaluate the meaning of the effects, he recommended that we consider how well the actual experimental conditions met the theoretical assumptions of the experimental design by examining descriptive frequency counts and logs. He went on to demonstrate with an example of a sensory discrimination experiment that the sensitivity of significance test is increased by increasing the number of cases or trials.

The second assumption of independence also gets violated in every instance where individuals included in a field experiment are nested in organizational units, such as patients nested in clinics or students nested in classrooms and schools, an often-unavoidable circumstance. Multilevel statistical models are now used in educational research to address nesting issues and longitudinal change questions; however, small numbers of cases in field experiments often test the limits of the accompanying estimation methods.

Matched sample quasi-experiments would likewise, in purist statistical terms, violate the second assumption. In empirical terms, it may be argued that matching independent cases on

a few characteristics does not create the same levels of equivalency (see Glass & Hopkins, 1984) or the same degrees of dependency as with the same cases measured on multiple occasions or with studies of paired twins or siblings and/or family members—conditions for which paired-samples statistics were designed. In any event, there are many ways in which dependencies result in correlated errors or in small numbers of cases that belong in nonequivalent comparison groups and/or cells within randomized field experiments. With violations in the assumptions of independence and statistical equivalence varying in different ways, decision making on statistical tests is complicated and far from an exact science. It is not well established how error rates affect various statistical significance tests under the variety of contingencies one faces under complex field research environments, nor which statistical tests have the most usable properties. The decisions are not easy for most laboratory experimentalists, let alone field researchers.

Given such conditions, scaffolding the statistical data on outcome effects (as obtained from field-based comparative experiments, whether RCTs or quasi-experiments) with ongoing, qualitative, and descriptive data adds depth, meaning, and clarity to findings. An historical recommendation on this point came from W. I. B. Beveridge (1957) in *The Art of the Scientific Investigation*, with reference to instances in early experimentation in the biological sciences. He specifically pointed to the need for methodically collected, descriptive information—such as detailed logs that document shifts in experimental, placebo (control or comparison setting), and environmental conditions—to reinforce and improve interpretability of results and avoid misleading conclusions. He used these words to warn us against overinterpreting statistical effects:

> The use of statistics does not lessen the necessity for using common sense in interpreting results. . . . Fallacy is especially likely to arise in dealing with field data in which there is a significant difference in two groups. This does not necessarily mean that the difference is caused by the factor which is under consideration because possibly there is some variable whose influence or importance has not been recognized. This is no mere academic possibility as shown in confusions in . . . many experiments with vaccination against tuberculosis, the common cold, and bovine mastitis. (p. 22).

In education as well, scholars have long asserted the fundamental assumptions of traditionally conceived, laboratory-style experimentation get grossly violated in the typical settings where most programs operate (see Campbell, 1981; Cronbach & Associates, 1980; Salomon, 1991) recommending before and after studies, and more systemic and mixed-method approaches. Drawing on that body of work, I proposed the use of ETMM designs with five guiding principles for their implementation (Chatterji, 2004, pp. 7-13):

- targeting a significant part of the life-span of a program for study in real-time environments
- incorporating distinct but self-contained phases of research that build on each other—an exploratory phase with formative goals aimed to help stabilize the "treatment" and understand the typical environment along with the comparison or "placebo" condition; and a confirmatory phase to formally examine effectiveness, using possible scaling-up to obtain sufficient sample sizes and ecological validity
- letting the data collection and analytic design be guided by the underlying logic of the program (the "program theory") and empirical information collected in situ on treatment versus control conditions and relevant environmental variables likely to influence, contaminate, or confound results over time

- delaying implementation of formal experiments to the confirmatory phase, preferably after causal questions can be sharply focused, treatment and/or control conditions better defined and environmental variables better documented, making them more amenable to direct observation, analysis, and interpretation
- combining use of qualitative and quantitative evidence in both phases to draw more comprehensive conclusions on how the program evolves and operates in real-time settings, its effectiveness, and the meaning and generalizability of the effects found.

The above principles are applicable to either academic or evaluation research carried out in field settings. When doing evaluations, a practical consequence of applying the above design principles is that the research can be more inclusive of stakeholders and oriented toward program development in Phase 1, the exploratory phase. A benefit that can accrue in terms of quality of confirmatory evidence in Phase 2 is that analytic designs and methods of observation and/or measurement can be refined to best fit the complexities of the program and its environment, as found in Phase 1.

The use of more long-term and phased "impact evaluation" designs that move from exploratory to more confirmatory phases has also been recommended for research on drug treatments. For screening evidence on cancer drug research, the National Institutes of Health (NIH) offers guidelines specifically outlining the value of three phase designs. Research and development efforts on new drugs, according to NIH, should dedicate Phase 1 to determine the best delivery methods and safe dosage levels—these studies may be labeled as "process monitoring studies" in evaluation terms; Phase 2 should examine whether the treatment, implemented optimally, yields the desired effect—that is, studies exploring preliminary process-to-outcome links; and Phase 3 should examine controlled comparative effects against existing remedies—or formal "impact evaluations." Carefully implemented RCTs would particularly apply in the third phase of effectiveness research on drugs, according to NIH recommendations.

Quite independently and speaking from the perspective of evaluating mental health programs for prison inmates, Wolff (2000, p. 107) expressed concerns with traditionally conceived RCTs in effectiveness research, seeking ways to improve the robustness of field studies. Her guidelines for "stylizing" traditional RCT designs for what she called "socially complex service interventions" included these recommendations:

- seeking and documenting that (patient and/or client) samples and environments are representative of real-world situations
- attending to valid observation and statistical documentation of treatment processes and unmeasured variables in the environment
- implementing a two-level RCT design that examines environmental variables at the first level and the patient and/or client factors at the second
- incorporating the use of qualitative methods along with RCT methods in achieving a better understanding of effects.

Are Studies From Medicine "the Gold Standard" in RCT Implementation?

Because of their early push for evidence-based practices and widespread use of randomized clinical trials, experimental studies in medicine have been held up as the "gold standard" for educational researchers to follow. In a 1999 conference hosted by the American Academy of the Arts and Sciences, influential scholars and government officials observed that because

the use of RCT designs in education was limited, the available evidence that could inform educational practice and policy lacked scientific credibility. The call then was for a direct transference of the standards of scientific research in medicine, defined narrowly based on the use of RCTs, to education.

As has been argued by several commentators since, programs in education, public health, and other social service sectors are rarely akin to tightly defined surgical procedures or pills. Rather, they comprise multiple, often-complex operational components that include changing human and social elements. As shown earlier, they operate in open, equally complex, and dynamic social systems. Such differences have been discounted or downplayed in the aforementioned federal recommendations. Regardless, the literature review in this article will now show that even with studies of well-defined medical treatments, an overreliance on RCTs alone can be problematic in cases where the medium of treatment and service delivery is socially mediated.

Some recent and visible medical studies have revealed empirical shortcomings of large-scale RCTs used alone. Specifically, research dealing with osteoporosis prevention with calcium supplements and arthritis medications (Cox-2 inhibitor studies) have yielded controversial and counterintuitive findings of long-term experiments—findings that have generated policy debates on drug safety (see Urquhart, 2005) because they are difficult to interpret and translate to medical practice. Concurrent documentation of the serious challenges faced by researchers in medicine and health in implementing RCTs—like those found in education—have added doubts to the utility of narrow, textbook-style experimentation with "socially complex" field interventions in health policy arenas, including medicine.

When compared with education, are the implementation challenges for RCTs in health and medicine really that different? And, is the quality of scientific evidence from medical and/or drug trials superior simply because of RCT use? Cautionary pointers from some commentators in medicine and health between the 1980s and 2000s are instructive. To conclude, this section will make specific reference to the Cox-2 inhibitor studies and the need for broader appraisal criteria for research evidence by highlighting the parallel issues in drug safety clinical trials and evidence-based medical practice and/or policy on the one hand, and educational research and evaluation on the other.

Sample Size and Outcome Measure Issues

Writing in the journal *Controlled Clinical Trials*, Boissel (1989) identified two main design issues that constrained most RCTs in medicine well over a decade ago that posed potential barriers to the practice of evidence-based medicine: that clinical trials were inadequately sized (with small numbers of patients) and that their outcome measures were inappropriate. As indicated earlier, small sample sizes constrain confident use of particular statistical tests and raise issues about sample representativeness and generalizability of results. Defensibility of outcome measures is, of course, key to capturing and interpreting effects in a meaningful way. To be sensitive to the intervention, the selected outcome measures must have substantive links to it (referred to as the property of content-based validity, a necessary part for ascertaining overall construct validity of measures; see American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), as well as other relevant psychometric properties, such as reliability of scores. Examining why evidence-based medicine was not the norm in nations such as France, the United States, Finland, and Holland at the time, Boissel pointed to a gap between medical practices and available scientific evidence on particular treatments and procedures.

Large-Scale Multisite RCT Issues

Others, such as Meinert (1980), similarly identified small numbers of participating patients in single-clinic studies, and advocated tryouts with multicenter RCTs to overcome this limitation in medical studies where the patient pool was likely to be small. Mienert's argument was that multicenter RCT studies would add numbers and ecological validity to the findings because the heterogeneity of site conditions would be more reflective of real differences in clinics across the nation. At the same time, he recommended that more research continue on the merits of multisite RCT methodology.

A counterargument to Mienert's recommendation would be that postselection biases and diverse environmental contingencies that cannot be successfully eliminated with randomization done at the start of an experiment multiply enormously when RCTs are implemented on a large scale—making the burden heavier for field researchers and evaluators in identifying confounders, mediators, and moderators once the study is under way. If such outside factors are ignored, findings become uninterpretable and confusing to researchers themselves as well as prospective consumers of the research evidence. In education, recommendations for multisite evaluations have been readily adopted and disseminated as the emphasis on scientific research through NCLB, and ESRA has become a national imperative. (For example, very large-scale, multiyear, multisite national evaluations were funded by the U.S. Department of Education with supplemental education and technology-based interventions.)

Generalizability and Internal Validity Trade-Offs in Small-Scale RCTs

Small-scale RCTs are not necessarily a panacea, unless there are alternate methodologies to bolster their limitations. In medicine, Davis (1994) identified generalizability issues with small-scale RCTs in studies with cholesterol reduction medication because of the limited ecological validity of very tightly controlled field experiments. Ecological validity speaks to the extent to which conditions of an experiment mirror typical environmental conditions where the treatment is expected to be implemented.

There is always a trade-off between internal and external validity of experiments, affected by design choices of researchers. The classic source that describes this design tension is the Campbell and Stanley (1963) book titled *Experimental and Quasi-Experimental Designs for Research*, where they delineated eight factors that could potentially threaten the internal validity, with four jeopardizing external validity of various experimental and quasi-experimental designs in their taxonomy. Internal validity issues prevent conclusive causal linkages between the treatment and outcome variables. External validity barriers affect generalizability of the findings to other cases, settings, and time frames of "known character." Campbell and Stanley stated: "Both types of criteria are obviously important, even though they are frequently at odds in that (research design) features increasing one may jeopardize the other" (p. 5, parenthesis added). Thus, even in the best of research circumstances the evidence from field research will have limitations linked to design actions and constraints—the central issue is whether researchers make an effort to be conscious of these factors, and if so, to what extent do they take steps to enhance the quality of evidence?

Issues Related to Treatment, Placebo, and Context Factors

Context-related interferences and vague treatments are not unique to education. A barrier to clean interpretation and scientific credence of findings of stand-alone, longitudinal RCTs concerns the oft-unknown administrative circumstances in hospitals and operational distinctions between a treatment program and the alternate or "placebo" condition. In medicine,

Klimt (1981) used examples of longitudinal RCTs in cardiovascular studies to point out that administrative and organizational factors in the larger context of the experiment affected design implementation, which in turn interfered with the meaningfulness and significance of results. In addition, he said that uneven adherence to treatments under experimental conditions and the inability to test what occurred under the placebo conditions remained a problem that confounded conclusions. Interaction with nonstudy drugs as environmental moderators was another problem.

All of the above design issues are further complicated by social and human elements in treatment administration and the environment in which cases are studied. Hospital and clinic-based studies in medicine are not vastly different from educational studies in this regard. For instance, doctors, nurses, hospital staff, family members, and patients themselves are often involved in administering prescribed treatments in given dosages, with the larger health care system and community offering an array of environmental conditions that mediate and/or moderate treatment delivery, and consequently the effects.

Issues Surrounding Units of Randomization

An additional design issue, related to the one just discussed, has to do with which sampling units are randomly assigned to treatment and control conditions in setting up the conditions of an experiment. This is a conceptual and a sample size issue that is irregularly implemented in medicine, education, and other fields. For instance, if a reading or math curriculum is being tested out by individual classroom teachers, it is classrooms that should be the units of the randomization process, with students nested within. To optimize randomization with an adequate sample size in this case, the units of focus would be classrooms, not students. With schoolwide initiatives, the appropriate units of randomization should be schools, and numbers of schools should be sufficient to claim equivalent status in treatment and control conditions on variables most likely to affect outcomes. In medical trials, if doctors in different clinics test out a medical procedure, it is the clinics that are the units of randomization—not the patients nested within. This design principle, is again, not well understood nor properly applied across fields of study.

Research Design Lessons from the Cox-2 Controversy

To sum up, what are the lessons from the Cox-2 studies? A central issue has been whether the statistical significance of findings on outcome studies of relative risk of harm with drugs like Vioxx, are ipso facto grounds for drug withdrawal from the marketplace and vice versa, when extraneous environmental factors, such as lifestyle choices and behaviors of patients, are excluded from designs. Alternate analyses that account for lifestyle factors of participants as additive or interactive factors have now yielded quantitative estimates of relative risks in the same statistical range as Cox-2 (Yusuf et al., 2004). Drug safety and policy analysts (Urquhart, 2005, p. 146) have thus recommended continuing "orderly . . . analyses" to understand better how a more comprehensive set of factors interact with other lifestyle indicators or mediate risks.

Another design issue discussed by the same author (Urquhart, 2005) that should have a familiar ring to readers of this article concerns the *N*—or number of patient participants needed in a trial that will optimize the statistical estimates of risk. A popular rule followed in drug safety studies is that with a homogeneous population of *N* exposed individuals who have no adverse reaction, one can be 95% confident that the incidence of risk over time will be 1 in *N*/3 per treatment cycle (Hanley & Lippman-Hand, 1983). Urquhart (2005) warned,

> The assumption of homogeneity, of course, is the weak link in the story, for patients differ by age, gender, co-morbidity, and the like. Thus, if . . . 10% of (drug-) exposed patients are truly at risk for the adverse reaction in question, then the "rule of 3" becomes the "rule of 30." (p. 146)

He went on to say that such issues can be obviated by larger and more costly trials and "by better prior learning" (p. 146). He cited a similar recommendation for more in-depth exploratory and confirmatory research in medical trials from Sheiner (1997), who differentiated between "learning" and "confirming" studies as necessary components of sound drug evaluation research.

Enhancing the Utility of Evidence from Quasi-Experiments and Nonexperimental Research Designs

In federal policy tools distinguishing sound effectiveness research from unsound (see, e.g., the What Works Clearinghouse standards in education at http://www.whatworks.ed.gov), evidence from quasi-experiments is considered acceptable but not as rigorous as RCTs. Nonexperimental designs generating quantitative, correlational evidence are not considered to lie within the realm of acceptability; that is, the latter are excluded from discussions of "strong" evidence altogether. Unqualified, such statements encourage erroneous generalizations, as much depends on considerations of aforementioned environmental factors, timing, and appropriateness of confirmatory analyses with respect to the developmental stage of a treatment program. Some clarifications will be made here on the utility of evidence generated from these two classes of research designs and how the evidence can be enhanced with complementary methods of data gathering and analysis, to allow for defensible causal inferences and generalizable links.

Like RCTs, quasi-experiments allow researchers to manipulate the intervention variable (the program) and, if other interfering factors are well controlled, can potentially yield causal evidence—or evidence to support a conclusion that a manipulated change in X causes a corresponding change in Y. The dual attributes of comparison and manipulation grant quasi-experiments a design status somewhat parallel to RCTs. However, quasi-experiments begin with a handicap with respect to creating the ideal conditions of independence, randomness, and equivalence of groups on preexisting characteristics governed by chance. A variety of design options are available to researchers choosing to go with quasi-experiments (Campbell & Stanley, 1963) to improve internal validity of the experiments; however, the strategies for establishing equivalence (such as matching cases in T and C groups) introduce their own limitations and frequently get jeopardized even further in the course of a study because of inhospitable field conditions. Under such conditions, systematic and ongoing documentation with multiple research methods can offer insights into the extent to which the evidence can be trusted and what the data comprehensively mean (see Chatterji, Kwon, & Sng, 2006, for a study that attempted this). In other words, acknowledging that they are initiated with design handicaps that can rarely be overcome or adequately corrected once an empirical study is in progress; the quality of evidence from quasi-experiments can still be improved by scaffolding the same with appropriate descriptive and qualitative evidence.

With non-experimental designs where a treatment and/or program variable cannot be manipulated (often called *ex post facto*, or *causal-comparative* designs), researchers must work with preexisting variables as the independent variables and examine their influences on desired outcomes. Because of the lack of ability of experimental manipulation, the evidence

in such studies is at best correlational—denoting associations among variables but disallowing definitive claims about causal links.

Some treatment programs and policy initiatives, however, simply do not lend themselves to randomized manipulation because of ethical reasons or practical and/or political barriers. For example, grade retention as an educational intervention can rarely be randomly assigned. Yet compelling evidence has accumulated over time with mostly ex post facto causal-comparative studies or quasi-experiments in education that point to its inefficacy over the long term (see, e.g., Holmes, 1989; Shepard & Smith, 1989). Such evidence has been largely ignored by politicians and policy makers under the current NCLB climate, where students are routinely held back in particular grade levels by policies intended to boost school averages on high-stakes state tests. In the health arena, the body of largely correlational research on smoking and cancer has led to very stable conclusions, suggesting causal connections without the need for randomized trials or quasi-experiments by manipulating the ethically undesirable intervention of cigarette smoking (see Raudenbush, 2002). Replicated correlational and explanatory evidence generated through well-done non-experimental research can thus be useful and should not be dismissed offhand as weak. In particular, causal path analytic approaches driven by strong theoretical rationales, with statistically controlled "exogenous" or "intervening" variable influences, have been recommended for confirmatory evaluations (see Reynolds, 2005). Adding the use of multilevel analytic models with appropriately selected and measured covariates at different levels, with scaffolding qualitative contextual data, can significantly improve the grade of evidence—making a compelling case as a whole for causality. In program evaluation contexts, phased implementations of non-experimental designs of the kind described can enhance the quality of correlational evidence when it is impossible to successfully manipulate interventions or control natural environments.

In sum, many compensatory mechanisms are available to enhance the grade of evidence with field-based effectiveness studies, starting with a systematic, phased approach to the research that is guided by a program's underlying logic and developmental stage. Some of these pertain to the initial design and setup of T and C groups to maximize internal validity (as shown in Campbell & Stanley, 1963); others address theory-driven, quantitative analytic approaches applicable to quasi-experimental and nonexperimental designs; others point to ways in which properly collected descriptive and qualitative data on environmental factors can add value as interpretive evidence by enhancing meaning and clarity of statistical effects found through comparative experiments.

Conclusions: Seeking the Best Grade of Research Evidence

Three conclusions can be drawn from the preceding literature review with regard to a "gold standard" in scientific inquiry and "strong" versus "weak" evidence tied to RCT use. First, RCTs (or any research design, for that matter), however foolproof in theory, get compromised to different degrees during the conduct of research under field conditions, generating grades of evidence. The design and implementation barriers are similar with complex programs and interventions across disciplines (medicine, health, psychology, and education). Overselling the case for RCTs, without adequate qualifications as to the grade of evidence actually obtained, may do a disservice to the cause for evidence-based practices by misleading consumers of research.

Second, RCTs are not only extraordinarily difficult to mount and carry out in field settings but also are never adequate by themselves for making generalized causal inferences on programs that are complex and field based. Because there are so many ways in which RCTs get undermined during field implementation, the evidence generated varies with respect to

degree of quality. The knowledgeable and astute researcher is conscious of relevant sources of error. Diligent researchers also make the effort to better study and understand how findings may be affected by various interferences in situ. That is, they make conscious design and analytic choices to enhance the quality of evidence from a field study by complementing RCTs with other, appropriate research methods. They evaluate not simply the statistical significance of effects but also the size, directionality, and meaningfulness of effects, given empirical information on program characteristics, alternate conditions, and environmental influences. More holistic documentation and appraisal of evidence adds clarity, meaning, and usability to the findings—aiding the cause for evidence-based practices.

Third, absolute distinctions between "strong" and "weak" research evidence and reliance on a "gold standard" for research designs are neither realistic nor useful from the perspective of utilization of research and evaluation results. What may be more useful are guidelines that help consumers sift through grades of evidence on program efficacy, or likely program efficacy, from classes of research designs with the potential to shed light on effectiveness questions with complex field-based programs. In evaluating grades of evidence, quasi-experimental and nonexperimental studies reinforced with relevant forms of alternate research evidence must be considered alongside RCT studies that utilize similar strategies. More work must thus be directed toward broader, if different, guidelines to identify trustworthy and scientific evidence from effectiveness research.

Moving Toward New Evidence Appraisal Criteria

As with extraction of different grades of usable oil products from crude petroleum, different research designs can potentially generate effectiveness evidence of varying degrees of quality with field-based effectiveness research, ranging from very high to very low grade evidence. However, the grade of evidence is also a function of the rigor with which studies can be and are executed under field conditions, after design choices have been made. Particular methodological and analytic choices made by researchers in response to field constraints and conditions they identify at different times during the conduct of a research or evaluation study are critical factors that determine the grade of evidence a study will yield.

A higher grade of evidence starts with the formulation of the very questions that guide an effectiveness study. The traditional question guiding RCTs (On average, does a new treatment or program have a statistically significant effect on desired outcomes in members of a target population, as compared to a control program?) is rather limited and less than ideal for obtaining a comprehensive and useful body of effectiveness evidence on the impact of complex social interventions. Questions must thus be broadened to illuminate the operational elements, developmental stage, and workings of a program in contrast with the alternate program(s), while documenting the environment where effects are found to be manifested.

To classify research evidence by grade, it may be time for the research and evaluation community to consider added criteria to progressively evaluate the evidence along a series of design dimensions that draw on the preceding literature. RCTs have important strengths for yielding causal evidence; however, they are inadequate by themselves for ascertaining the impact of complex social interventions. The grade of evidence from RCTs can be improved with qualitative and descriptive data gathering on relevant factors, after a defensible application of random assignment of cases to T and C conditions and appropriate outcome measures selected. A further improvement occurs if the randomized experiment is delayed until formative "learning" studies are completed to help a new program mature and function as a coherent entity in actual environments. Confirmatory studies, utilizing appropriate analytic models, should ideally be scaled-up efforts that utilize what is learned from small-scale program testing in earlier phases.

Notes

1. See, for example, Mathematica Policy Research, Inc., and Decision Information Resources, Inc. (2003).

2. The Study Design and Implementation Assessment Device (Study DIAD) is a screening instrument intended to identify rigorously conducted effectiveness research on educational interventions and can be found on the What Works Clearinghouse (http://www.whatworks.ed.gov) Web site established by the Institute for Education Sciences.

3. As in my earlier articles on this topic, I broadly classify social science inquiry under two broad categories: academic and evaluation research. Academic research is initiated by independent scholars interested in particular questions or scientific issues, geared mainly toward formal theory development and expansion of knowledge on a phenomenon, and aimed toward specialized audiences in academe and the professional field. Evaluation research typically emerges in response to client and/or stakeholder information needs, is typically motivated by social action or betterment ideals, and aimed mainly at informing decisions and actions of policy makers, sponsors, field practitioners, and stakeholders, with knowledge production coming after. Although academic research may be conducted in laboratory, simulated, or field settings, evaluation research must be conducted in real-time conditions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18-20.

Beveridge, W. I. B. (1957). *The art of scientific investigation*. New York: Norton.

Boissel, J. (1989). The impact of randomized clinical trials on medical practices. *Controlled Clinical Trials, 10*(4), 120-134.

Campbell, D. T. (1981). Introduction: Getting ready for the experimenting society. In L. Saxe & M. Fine (Eds.), *Social experiments: Methods for design and evaluation* (pp. 13-18). Beverly Hills, CA: Sage.

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Chatterji, M. (2004). Evidence of "what works": An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher, 33*(9), 1-13. (Reprinted: Chatterji, M. (2005). Evidence of "what works": An argument for extended-term mixed method (ETMM) evaluation designs. *Educational Researcher, 34*(6), 13-24.

Chatterji, M., Kwon, Y. A., & Sng, C. (2006). Gathering evidence on an after-school supplemental instruction program: Design challenges, lessons, and early findings in light of NCLB. *Educational Policy Analysis Archives, 14*(12), Retrieved on September 6, 2006, from http://epaa.asu.edu/epaa/v14n12/.

Cronbach, L. J., & Associates. (1980). *Toward reform in program evaluation.* San Francisco: Jossey-Bass.

Davis, C. E. (1994). Generalizing from clinical trials. *Controlled Clinical Trials, 15*(1), 11-14.

Education Sciences Reform Act of 2002, H. R. 3801, 107th Cong. (2002).

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher, 31*(8), 4-14.

Fisher, R. A. (1960). *The design of experiments.* New York: Hafner Publishing.

Glass, G. V., & Hopkins, K. D. (1984). *Statistical methods in education and psychology.* Englewood Cliffs, NJ: Prentice Hall.

Hanley, J. A., & Lippman-Hand, A. (1983). If nothing goes wrong is everything all right? Interpreting zero numerators. *Journal of the American Medical Association*, *249*, 1743-1745.

Holmes, C. T. (1989). Grade-level retention effects: A meta-analysis of research studies. In L. A. Shephard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16-33). London: Falmer.

Klimt, C. R. (1981). The conduct and principles of randomized clinical trials. *Controlled Clinical Trials*, *1*(4), 283-293.

Hsieh, P., Levin, J. R., Acee, T., Chung, W., Hsieh, Y., Kim, H., et al. (2005). Is educational intervention research on the decline? *Journal of Educational Psychology, 97*(4), 523-529.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational psychology's credibility gaps? *Issues in Education: Contributions from Educational Psychology*, *5*, 177-229.

Mathematica Policy Research, Inc., & Decision Information Resources, Inc. (2003). *When schools stay open late: The national evaluation of the 21st century community learning centers program.* Jessup, MD: U.S. Department of Education, ED Pubs.

Meinert, C. L. (1980). Toward more definitive clinical trials. *Controlled Clinical Trials*, *1*(3), 249-262.

No Child Left Behind Act of 2001, P.L. 107-110, 115 Stat. 1425 (2002).

Raudenbush, S. (2002, February 6). *Scientifically-based research*. Paper presented at the U.S. Department of Education "Use of Scientifically Based Research in Education" Working Group Conference, Washington, DC.

Reynolds, A. J. (2005). Confirmatory program evaluation: Applications to early childhood interventions. *Teachers College Record*, *107*(10), 2401-2425.

Saloman, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher*, *20*(6), 10-18.

Schoenfield, A. H. (2006). What doesn't work: The challenge and failure of the what works clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher, 35*(2), 13-21.

Sheiner, L. B. (1997). Learning versus confirming in clinical drug development. *Clinical Pharmacology and Therapeutics*, *61*, 275-291.

Shepard, L. A., & Smith, M. L. (1989). Synthesis of research on grade retention. *Educational Researcher*, *47*(8), 84-88.

Urquhart, J. (2005). Some key points emerging from the COX-2 controversy. *Pharmacoepidemiology and Drug Safety*, *14*(3), 145-147.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide.* Washington, DC: Institute for Education Sciences.

Wolff, N. (2000). Using randomized controlled trials to evaluate socially-complex services: Problems, challenges, and recommendations. *Journal of Mental Health Policy and Economics, 3*, 97-109.

Yusuf, S., Hawken S., Ounpuu, X., Duns, T., Avezum, A., Lanas, F., et al. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *Lancet, 364*, 937-952.