



Evidence on “What Works”: An Argument for Extended-Term Mixed-Method (ETMM) Evaluation Designs

by Madhabi Chatterji

Federal policy tools for gathering evidence on “What Works” in education, such as the What Works Clearinghouse’s (WWC) standards, emphasize randomized field trials as the preferred method for generating scientific evidence on the effectiveness of educational programs. This article argues instead for extended-term mixed-method (ETMM) designs. Emphasizing the need to consider temporal factors in gaining thorough understandings of programs as they take hold in organizational or community settings, the article asserts that formal study of contextual and site-specific variables with multiple research methods is a necessary prerequisite to designing sound field experiments for making generalized causal inferences. A theoretical rationale and five guiding principles for ETMM designs are presented, with suggested revisions to the WWC’s standards.

The What Works Clearinghouse (WWC) was recently charged by the Institute of Education Sciences ¹(IES) with collecting, appraising, and reporting on the “strength and nature of scientific evidence on the effectiveness of education programs, products, and practices” (see <http://www.ed.gov/offices/OERI/whatworks/>). In response, the WWC developed and recently released a set of standards for selecting empirical studies yielding research-based evidence on effective educational programs (Valentine & Cooper, 2003).

Consistent with the language and references to scientifically based research in the No Child Left Behind Act of 2001, the language in the WWC standards also refers to testing of hypotheses using experimental designs, with a preference for randomized field trials for judging effectiveness of educational initiatives (Eisenhart & Towne, 2003). The larger mission of the WWC appears to follow in the traditions set by the Cochrane Collaboration in medicine and the Campbell Collaboration in education, where the aim is to collect and synthesize results from a series of topically-related experimental studies with meta-analyses, permitting more dependable inferences to be drawn about particular interventions than is possible with any single study.

This article has been reprinted to correct copy-editing and author errors that appeared in the original version, published in the December 2004 issue of *Educational Researcher*. For citations and quotations, please refer to this version of the article.

The WWC standards are formulated as four “general”, eight “composite”, and a number of more specific questions on a screening instrument, titled the Study Design and Implementation Assessment Device (*Study DIAD*) (See w-w-c.org). Albeit labeled as drafts and preceded with a caveat that acknowledges other methods, the current WWC standards unambiguously recognize comparative experimental designs as the predominant, if not the sole, approach for making definitive appraisals of “what works” in education. In the rush to emphasize generalized causal inferences on educational initiatives, the developers of the WWC standards end up endorsing a single research method to the exclusion of others. In doing so, they not only overlook established knowledge and theory on sound evaluation design; they also ignore critical realities about social, organizational, and policy environments in which education programs and interventions reside.

Thoughtful protests from renowned leaders of the American Evaluation Association (AEA) ²and the American Educational Research Association (AERA) (see St. Pierre, 2002; Berliner, 2002; Erikson & Guitierrez, 2002; Feuer, Towne, & Shavelson, 2002; Pellegrino & Goldman, 2002) are testament to the fact that the notions of “science”, as applicable to gathering research-based evidence on school interventions and programs, have been fundamentally mischaracterized in federal documents stemming from the NCLB legislation, and now, the WWC guidelines reflected in the *Study DIAD*. That the protests surfaced as the original proposal for developing the *Study DIAD* was supported by distinguished experimental scholars of the AEA (Valentine, personal communication), suggests a continuing need for dialogue and resolution of research design issues among members of the academic community. The mission of evaluation researchers today has broadened to an extent where “impact evaluations” involving generalized causal inferencing are just one of many models that are viewed as useful in addressing critical social problems.

Before the WWC standards become consolidated as edicts for research and evaluation practice, however, a close examination of the *Study DIAD* is warranted. The consonance of the WWC standards with agreed-upon methodological principles for field-based studies in education and social sciences, must be critically evaluated. As a part of the deliberations, alternate research designs better equipped to support generalized causal inferencing with field-based interventions, need consideration. Definitions of “science” and best research practice that emerge must accurately reflect the progress made to date in evaluation theory. Limited and one-sided criteria in federal policy tools, carrying enormous potential to influence school audiences, consumers, and funders of

research, must be revised until they reflect principles for more complete and effectual research designs.

Purpose

The primary purpose of this article is to offer a counter-methodology and argument for designing rigorous evaluations of educational programs, policy initiatives, products, services and interventions (referred to hereafter as *programs*), taking into account the fundamental characteristics of sound evaluation research, and the complex organizational and community environments in which educational programs evolve and typically operate. Instead of a singular dependence on experimental designs, this article calls for the use of *extended-term mixed-method* (ETMM) designs for making generalized causal inferences on “What Works”. ETMM designs follow life-spans of individual programs/policy initiatives within particular environments, employing appropriate descriptive research methods in the early stages of program adoption and implementation followed by timely and judicious implementation of experimental designs at a subsequent stage.

Underlying the argument for ETMM designs is the basic premise that there is a temporal factor to be considered in gaining thorough understandings of programs as they develop and take hold in organizational or community settings. Reasonable questions to ask about a particular program at a particular time, and methods best applied to answer them are thus predicated on the developmental stage of the program as it operates at a given site. In-depth and often site-specific studies of context variables, along with systematic examinations of program inputs and processes as potential moderators and intervening factors, are a *necessary prerequisite* to both designing and implementing sound field experiments geared towards answering causal questions on program impact. Further, very tightly conceived but de-contextualized experiments, following in the research traditions of laboratory experimentation, are weak research designs for studying educational programs in field settings. In the same vein, prematurely implemented experimental designs do not lead to improved understandings of “what works”. Rather, what often results is an atheoretical, poorly conceptualized, “black box” evaluation (see Rossi, Freeman, & Lipsey, 1999, p. 154 for a definition) where little is unveiled as to the reasons and conditions under which a program worked (if indeed desired outcomes were manifested), or the causes for its apparent failure (in the event outcomes could not be documented). External validity and replicability issues—critical for program expansion and dissemination—remain unresolved.

The point of this article is *not* to undermine experimental methods or the intent of the WWC’s *Study DIAD* or the NCLB legislation on gathering sound research-based evidence, at all. Rather, the article is a call for more *contextually-grounded* and *suitably-timed* implementation of experimental designs, the latter defined more broadly than “randomized control group trials”. It is a call for a *broader set of guidelines* (whether disseminated through federal policy tools such as the *Study DIAD* or otherwise) that will help both engender and guide the selection of field studies of more defensible quality than presently offered by the WWC standards. To these ends, the article identifies oversights and limitations in the present *Study DIAD*, drawing on academic discussions from the recent past.

In conceptualizing research in schools to determine “what works”, this article specifically speaks to the utility of three broad sets of ideas: scientific realism as opposed logical positivism as the epistemological foundation (House, 1991); context-sensitive and long-term systemic designs (Salomon, 1991; Stufflebeam, 1983; Stufflebeam, 2003); and the use of multiple research methods (Greene & Caracelli, 1997) guided by the purpose and evolutionary stage of the program. Acknowledging that the undergirding ideas for ETMM studies have been evident in the evaluation literature for at least two decades (see Cronbach & Associates, 1980; Campbell, 1981), this article undertakes the task of synthesizing the concepts into a coherent set of guiding principles. The claim is that ETMM designs, as described, encapsulate established methodological needs best; other approaches fall short and will compromise the quality of research evidence on field-based programs.

Embedded in the article, is a second but equally important purpose. The bias towards particular research methods in NCLB’s language on gathering scientific evidence, has led to legitimate concerns among some researchers on the types of research likely to enjoy federal support in future (see St. Pierre, 2002; Erikson & Guterrez, 2002). Keeping aside the different value positions and orientations of researchers/scholars, there is a current need to raise awareness levels held by policy-makers and lay consumers of research, as to what counts as effective field research in education. Given the unprecedented NCLB legislation, the need to change conventional and narrow notions of good research practice as only randomized trials is particularly high among prospective funding agencies. It is more important today than ever to educate and convince both federal and non-government funders on how best to recognize studies likely to generate sound field-based evidence on “what works”. It is through continuing efforts at discussion, education, and demonstration that the necessary support for ETMM-type studies can be obtained. This article begins that effort.

Lastly, explicit federal policy tools such as the WWC’s *Study DIAD* inevitably influence views of the public and the media on the meaning of educational science. They also shape values, understandings, and practices of those who either adopt and implement new school programs, or develop and seek clients for their products and services.³ The definition of “scientific evidence”, as given in federal policy documents must thus be examined closely, parsed, tested, and as necessary, revised, before their use becomes widespread and it is too late to alter course. One of the aims of this article, is thus to educate the broader community, including political leaders who influence NCLB-type laws and policy on evidence-based practices.

To make the case for extended-term mixed-method (ETMM) designs, the article begins by drawing on cumulative knowledge from the fields of social science and evaluation research and the author’s own experiences as an evaluation researcher over several years. Particular illustrative points refer to a recently completed study of a supplemental instructional program in a New York City public school.

Why ETMM Designs should be Preferred Over Experimental Methods

Several writers have eloquently made the case that the choice of method in scientific research should be governed by questions and purposes, rather than by ideological stances of scientists (see

for example, Feuer, Towne, & Shavelson, 2002). In addition, evaluation research, directly pertinent to evidence-based school practices, is fundamentally different from academic research with respect to purpose, audience, and conditions of work (Chatterji, 2002). In designing evaluations, the most useful theoretical ideas attend to the complexities of a program's "life" as well as its milieu. The first section of this article gives reasons as to why textbook assumptions regarding experimental designs do not translate well to evaluation settings and why field conditions make it necessary for researchers to draw from a broader set of research strategies and tools of inquiry.

Barriers to Implementing Textbook-Style Field Experiments

Per the literature, the simplest "true" experiment involves a two-group design and two variables—one independent variable (the IV, typically the treatment or intervention that is manipulated by the researcher) and one dependent variable (the DV, the expected outcome). In such a design, one group of subjects is randomly assigned to the treatment condition while the other (the control group) receives an alternate or no treatment at all. This design has been historically recommended as one of the better design strategies for addressing questions about cause and effect, because it upholds the principles of control, randomization, and comparison (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002; Spector, 1981). In a perfect experiment, the treatment variable is manipulated, subjects are randomly assigned, thus equalizing pre-existing inter-subject differences, and all other variables—extraneous to the experiment or potential confounders—are held constant, thereby permitting a conclusive link to be drawn between the treatment IV and the outcome variables DV.

A number of variations to the basic experimental design are possible, such as, multiple group, longitudinal designs or split plot designs. When random assignment cannot be done, quasi-experiments, in the form of matched group comparisons, interrupted time series designs, or regression discontinuity methods, are recommended as alternatives for answering cause-effect questions about interventions (Shadish, Cook, & Campbell, 2002).

The various experimental designs involve trade-offs among the principles of comparison, manipulation, and randomization (Spector, 1981). To different degrees, such trade-offs threaten internal validity of the experiment, preventing conclusive causal linkages to be made between the treatment and outcome variables, or to the external validity (generalizability) of the findings. Highly controlled experiments reduce external validity because they create unrealistic laboratory-like conditions that cannot be replicated in actual settings where an intervention is eventually implemented. More loose controls, on the other hand, diminish internal validity, permitting inferences only about the "gross effects" of the intervention⁴ rather than its "net effects" (Rossi, Freeman, & Lipsey, 1999).

A researcher's choice of a particular experimental design is contingent on availability of time, resources, subjects, and other practical constraints. Textbook threats to executing sound field experiments are documented as subject selection biases, differential history of subjects, problems with outcome measures (such as poor validity, reliability, or instrument reactivity), subject at-

trition, non-representative samples, and poor operationalization or lack of treatment fidelity in the experiment.

In educational evaluations, experimental designs are recommended when a study of program impact is undertaken (Rossi, Freeman, & Lipsey, 1999). However, the threats to conducting sound field experiments are complicated many times over by a large number of additional factors inherent in a program's environment. Nevertheless, when designing impact evaluations, particular interventions, programs or policy initiatives implemented in schools are viewed as "treatments", and the desired student outcome(s), as given by the program's theory, serve as the DVs.

How do evaluation conditions violate assumptions of textbook-style field experiments? Because forces outside the control of researchers often initiate educational programs (for example, a state-legislated Voucher Plan, a federally sponsored standards-based mathematics reform curriculum, a reading program adopted by a city/district system, a state legislative action on class-size reduction), the IV can rarely be locally manipulated by researchers, as expected in textbook descriptions of experimental methods. In addition, the IV is rarely, if ever, a single, discrete, narrowly scripted and easily identifiable condition in field settings.

Consider the example of class-size studies. The Tennessee STAR (Student Achievement Ratio) experiment was an impact study of a statewide demonstration project on class size. It involved longitudinal, randomized field trials in a large number of elementary schools. Such conditions are rare in educational studies, and that work has yielded fairly robust conclusions. However, here too, it was state sponsors/stakeholders, rather than researchers, who manipulated the IV as well as funded and authorized the study. In most studies of "class-size reduction", the definition of the IV becomes complicated by the presence or absence of volunteers, para-professionals, teacher interns, or other teachers, who along with the regular teacher in a classroom, provide different degrees of instructional support to a given number of students. This alters the ratio of instructional staff to pupils and the operational definition of the field "treatment" condition. Often treatment conditions may vary across classrooms and schools studied, and sometimes, even in the same classroom at different times of the school day. Without adequate documentation of these *qualitative variabilities*, the effects are hard to interpret, let alone replicate. Under the typically non-sterile and complex field conditions, thus, the approach to determine whether or not a program "worked" has to be significantly altered. It follows that the criteria for judging the quality of studies yielding evidence on programs must also change.

"What Works" Questions and Evaluation Research

Educational researchers engage in two fairly distinct kinds of inquiry on educational phenomena: academic research and evaluation research. This distinction was made during a recent analysis of a body of empirical research on standards-based educational reforms (Chatterji, 2002) and is useful in guiding the present discussion as well. "What Works" questions arising out of practice and policy environments, fall unquestionably in the category of evaluation research, and must thus be pursued with methods that suit field conditions.

In the typical case, independent scholars interested in particular questions or scientific issues, initiate academic research. Aca-

demographic inquiry is geared mainly towards formal theory-development and expansion of knowledge on a phenomenon. Further, such research is aimed towards specialized audiences in academe and the professional field. Basic research of the type conducted by cognitive psychologists or laboratory scientists, where narrowly defined and tightly scripted treatments are deliberately manipulated in given “dosages” by researchers to examine effects on a DV, are examples of the experimental method applied in an academically driven research program. Likewise, ethnographic studies of organizational cultures conducted by social anthropologists, when initiated by researchers themselves, are other examples of academic research.

By contrast, evaluation research typically emerges in response to needs of individuals who approach the research community with their information needs, and who are typically motivated by social action or betterment ideals (Henry & Mark, 2003). In its pure form, evaluation research is aimed first towards informing decisions and actions of policy-makers, sponsors, field-practitioners, and stakeholders of educational programs (see Joint Committee on Standards for Educational Evaluation, 1994). New knowledge production may often ensue from evaluation research, but takes second place in the effort.

While the same researchers may conduct both types of work and research questions and methods may overlap, academic research is first and foremost, “conclusion-oriented”, while evaluation research tends to lean purposefully first towards “decision-oriented” inquiry (terms taken from Cronbach & Suppes, 1969). And, while the former may occur in tightly controlled laboratory settings where a researcher has the authority to institute controls at will; the latter must occur in real-time environments where the object of study may be a demonstration or pilot program, or a program at a more advanced stage of implementation at one or more site(s). As a consequence, the IV or “treatment” variable may assume different qualitative characteristics as it permeates down and across different levels of the organizational or social system in which the program is housed. Without systematic study of these qualitative differences *in context*, thus, cause-effect questions are difficult, if not impossible, to answer.

By definition and consensus in the profession, then, sound evaluation research should occur not only with rigorous application of appropriate research methods, whether qualitative or quantitative (see Accuracy Standards 1-12, Joint Committee on Standards for Educational Evaluation, 1994), but also with a sense of responsibility and service towards prospective users of the information (see Utility Standards 1-8, Joint Committee on Standards for Educational Evaluation; Propriety Standards on Service Orientation). Identification of information needs of stakeholders; documentation/analysis of the program’s context; unraveling of a program’s underlying theory to illuminate key variables and their linkages; examining the consistency and authenticity of a program’s delivery processes against the program’s theory; providing stakeholders with timely information to shape and improve program delivery while in early stages of operation; educating stakeholders on a program’s readiness for a summative, impact study; and conducting timely, meaningful, and appropriately controlled impact analyses, are only a few hallmarks of well-conducted evaluations that are overlooked in the present WWC and NCLB criteria for scientific evidence (see Bickman, 2000; Bernstein, Whitsett, & Mohan, 2002; Cook, 2002; Joint

Committee on Standards for Educational Evaluation; Smith, 1989; Suchman, 1967; Weiss, 2000).

Within- and Cross-site Realities: A Program’s Life and Potential Moderators/Mediators

Outside the main (treatment/program) variable of interest in a study, psychologists often refer to other predictors that affect outcomes as *moderators*, and to factors that intervene as *mediators* (Baron & Kenny, 1986; Jaccard & Turrissi, 2003). This nomenclature will be used here to facilitate communication.

A consequence of the contextual realities in organization or social systems where education programs are typically implemented, is that the number of moderators is large and subject to change over time. Likewise, the treatment configuration itself may change and evolve as a program settles at given sites, inevitably conditioned by various factors in the environment. Context, input, process variables in a multi-level educational system (Stufflebeam, 1983, 2003) are potential moderators that could interact with and mediate effects on outcomes in numerous ways. Cronbach’s (1975) widely cited “hall of mirrors” argument raised consciousness of researchers to such contingencies. Some conditions may generalize to all target sites; others may be site-specific. Some may be transient; others may last through the lifetime of an intervention.

In the first year of implementation, for example, a legislative action on class-size reduction may vary from school to school, mediated by resources (an input variable) that individual schools are able to allocate. In addition, the achievement effects of a legislative action on class-size reduction at a particular grade level could vary in different subpopulations served by a school (a moderating context variable), such as Limited English Proficient and native speakers of English. Before evidence of effects can be sought with more targeted quantitative designs, qualitative and exploratory inquiry is necessary to pin down, select, and better understand the most important moderating and mediating variables within and across implementation sites. Dedication of time and resources to formally study the most critical variables *in situ*, thus, is standard practice for the prudent researcher.

Theoretical Base and Definition of an Alternate Design: The ETMM Approach

Given the contingencies just raised, what would be the defining characteristics of a useful, alternate research approach for making generalized causal inferences on educational programs? What theoretical conceptions could such an approach draw from? This section now makes reference to relevant theory, and offers defining criteria.

House (1991) forwarded the view of “scientific realism”, pointing to its utility in the conduct of educational evaluations and as a means for investigating causal structures in complex, multi-level, open systems where educational programs are typically found. In such environments, he recognized that multiple causal agents typically influence outcomes. Presenting scientific realism as an alternate conception of science, House opposed the “standard” view of logical positivism that reigned in psychological circles in the 20th century, heavily influencing traditions of educational research and evaluation prevalent through the 1980s.

Elements of positivistic thinking that are particularly limiting include the notions that causality is based on regular chains of contingent events (if x, then y); that laws and theories are hypothetical-

deductive systems; that predictability is the ultimate test of theory; and that the aim of all empirical science is prediction. Fundamental to the scientific realist's understanding of causality, in contrast, is the dedicated study of environmental conditions, along with substantively-guided causal interpretation and explanation. Rejecting cause-effect chains of events as construing social reality too simply, House advocated the scientific realist's approach, in which social researchers would "track variability and irregularity of events . . . , describing programs and their outcomes so that influences can be registered and . . . causal entities and their interactions be understood (House, 1991, p. 8). Elsewhere, House also pointed out that a search for sterile facts (evidence) on programs, divorced from value-orientations of society—including values of researchers themselves—is not an achievable end in most research and evaluation contexts.

Along similar lines, Salomon (1991) distinguished between the usefulness of "analytic" and "systemic" approaches to educational research. He characterized analytic approaches as the isolated study of discretely defined variables from complex educational phenomena, with the unrealistic assumption that all else will remain unchanged. He endorsed systemic designs for capturing the richness of events and actions in complex social environments, such as classrooms, focusing on a study of patterns among variables, and recognizing the inter-dependence, inseparability and transactional relations among elements. In the end, Salomon recommended that both approaches, analytic and systemic, be employed by researchers to *complement one another*, saying that analytic approaches might be the best way to test specific theory-driven causal hypothesis in certain circumstances. In ETMM designs advanced here, for instance, analytic designs could be incorporated in a social scientist's larger research plan once complex environments are better understood.

With respect to methodological tools, Greene and Caracelli (1997) summarized applications of mixed-method approaches in research and evaluation practice, citing triangulation (establishing construct validity through convergence in findings from multiple studies), complementarity (combining methods to obtain a fuller picture of a construct) and expansion (using more than one method to obtain a fuller picture of a program), as the main driving forces. A pragmatic stance to using mixed methods would entail a choice of methods guided by the questions and issues surrounding a particular study, independent of philosophical differences associated with quantitative-qualitative paradigm wars within the research community.

Earlier and more importantly, Cronbach and a multi-disciplinary team of associates wrote a book, *Toward Reform in Program Evaluation* (1980), calling for reform in program evaluation in the published proceedings of the Stanford Evaluation Consortium dated 1973–1979. Offering 95 "theses" to argue for changes in field research and evaluation practices, these authors acknowledged that, ". . . Time and again, political passion has been the driving spirit behind a call for rational analysis" (p. 4). Today's push for research-based evidence by the federal government actually provides new data to support this particular thesis. It is also interesting that *Toward Reform in Program Evaluation* emerged in the context of debates in the 1970s on the usefulness of randomized designs in social experiments!

Weighing complex issues in the conduct of field research, including hypothesis-testing under changing rather than fixed

conditions, configuration of suitable designs for sampling and scaling-up of multi-site interventions, and the political feasibility and ethics of withholding treatment to some in true randomized experiments, Cronbach and Associates (1980) argued for a "before-and-after study . . . beginning with development work in a few villages and increasing the sample only after pointed, significant questions arise that a small sample cannot answer" (p. 271). Two of Cronbach and Associates' (1980) organizing theses, in particular, speak to the rationale for ETMM designs presented in this article, namely:

A good evaluative question invites a differentiated answer, instead of leaving the program plan, the delivery of the program, and the response of clients as *unexamined elements within a closed black box*" (#37, p. 5, emphases added); and "Precise assessment of outcomes is sensible only *after thorough pilot work has pinned down a highly appropriate form* for an innovation under test" (#40, p. 6, emphases added).

Another major commentary published concurrently with that of Cronbach et al (1980) was Donald Campbell's introduction to Saxe and Fine's (1981) textbook on social experimentation. In it, Campbell articulated the ideal of an "experimenting society" (Saxe & Fine, 1981, p.14), a society that seeks solutions to recurring social problems through systematic field research and experimentation. He acknowledged therein that research inevitably becomes a political process when social experiments involve testing of governmental policies. Campbell recommended that researchers should acknowledge such environmental realities in their designs, not viewing them as limitations, but as factors that authenticate the research process when modeled appropriately. An advocate of quasi-experimental designs, Campbell concluded that use of *multiple* methods, including opinion surveys, and *multiple* social indicators, would in fact add validity to results of field studies despite adding a research burden to scientists. It should be noted that Campbell's work was published before the advent of qualitative methods as additional tools of inquiry in the social sciences and education.

To summarize, then, theorists have long recognized that comparative experiments *are insufficient* by themselves to garner the best evidence of "what works", however broadly researchers define them. A pragmatic and productive design solution is the Extended-Term Mixed-Method (ETMM) approach offered here, grounded in the aforementioned theoretical rationale and defined by the following characteristics:

- Use of a long-term research plan, deliberately tracking the course of a program or intervention over relevant parts of its life with formative and summative studies
- Use of systemic, contextually-grounded studies in early phases followed by more sharpened, analytic experimental/quasi-experimental studies in later phases of the research
- Deliberate study and documentation of environmental variables as a component of the research plan
- Combined use of more than one research method, uncovering of patterns and deepening understandings of relationships and causality
- Explanation of causality based on both empirical and substantive knowledge gained on the program *and* its setting.

Guiding Principles for Extended-Term Mixed-Method (ETMM) Evaluation Designs

ETMM designs could potentially guide field-based research on any new idea or intervention, whether categorized as *academic* or *evaluation research*. This article will specifically speak to five principles that might guide educational evaluations. The principles build on one another and serve overlapping objectives from a research design perspective.

Principle 1. ETMM evaluation designs employ a long-term time-line that targets a significant part of the life-span of a program at given site(s) for systematic study.

The life of a program in particular organizational or community contexts can be temporally charted from the time of its adoption, to early implementation and pilot-tryout stages, to the culminating, full-scale operational phase at the program site(s). In the planning and execution of an evaluation to determine whether or not the program “works”, a sufficient part of its life span should be targeted for formal study.

For clients, formal data gathering in the early life of a program could serve several different functions, including: needs assessments to guide initial program design and planning; context studies to promote understandings of systemic, social and political forces that affect program delivery, management, and outcomes; process monitoring studies to modify/tighten service delivery processes; resource allocation or cost feasibility studies to appraise program inputs; capacity needs studies to design effective program delivery components; preliminary process-outcome relationship studies to test the program theory at new sites; or program evaluability assessments to advise clients on a program’s readiness for formal evaluations.

For researchers, systematic investigations conducted during the early life of a program carry the important design benefit of providing much-needed insights on the dynamic nature of programs and the complexities inherent in their immediate and larger environments. They will also inform researchers on a program’s underpinning theory and reveal ways in which a program departs from that theoretical conception in practice (see Principle 2, elaborated next). Equalizing inter-subject differences with random assignment or matching is really only a small step in controlling for co-contaminants that may confound or interfere with understandings whether a program works in field settings. Localized co-contingencies and moderators (alternate causal factors) that may influence or mediate outcomes, and remain otherwise obscure or misunderstood by researchers, can become visible through early studies. Such knowledge can then support design decisions for later experimentation, such as, in identifying variables that potentially interact with the treatment condition, and decisions on critical interactions to test.

Initial studies also help researchers properly time the impact study based on empirically-grounded answers to the question: Is the program delivering services per theory and ready to be evaluated for effects? (Often called an evaluability assessment). Assuming that the program is not abruptly discontinued for reasons outside the researchers’ control, a longer term design strategy permits execution of experimental methods in ways that lead to more informed conclusions on program effects.

Principle 2. ETMM designs are guided by a program’s theory and empirically based understandings of environmental, systemic and site-specific factors that could potentially influence program outcomes.

All programs have an underlying logic, or a set of explicit and often, implicit assumptions that suggest how the desired outcomes should be affected by variables in their context, as well as by program inputs and processes. The underpinning logic model represents the “program theory” (Bickman, 2000; Suchman, 1967). The program theory is most obvious if there are clearly stated program goals, such as goals in a Title 1 School-wide Project, and documentation is available in published project descriptions, manuals, materials, professional development systems, or other program products and procedures, on how the “ideal” program would look and function. When less evident, developers, leaders, program managers, and delivery personnel are useful “key informants” on a program’s theory, and should be tapped for this information.

Program theory analysis often overlaps in purpose with program context studies (see Principle 1). They are also best conducted in the early stages of a program’s life. One result of a program theory analysis is that it generates logic models portraying expected causal links between critical program resources and processes, or between major processes and outcomes. Another result is a better understanding of extraneous and confounding variables in the open social and organizational system.

Once the program theory is charted, it sheds light on multiple program facets and linkages, and can guide decisions on which links to examine (see Weiss, 2000) in studies conducted at different stages of a program’s evolution. To evaluate authenticity of program delivery during process evaluations, for example, comparative studies can be designed to examine whether processes actually observed fit the processes that should be occurring “in theory”. The directions in which the observed processes influence desired outcomes may also be examined.

With ETMM designs, the qualitative characteristics of the “treatment” can be documented and evaluated against theory in real-time, as conditioned by site-specific factors. Some questions that can be attacked, for example, are: How do the program managers view and implement the program components? How do the program staff view and implement the services? Do program recipients respond to the services as expected? To what extent are the overall program operations consistent with theory? What systemic factors directly and indirectly impinge on targeted outcomes? Such knowledge can help prevent the design of very narrowly-conceived experiments in later stages, that overlook critical program components as a part of the “treatment” condition—leading to poor external validity or non-reproducible effects at other sites.

Principle 3. ETMM designs deliberately incorporate formative and summative evaluation phases in the overall research plan, with at least one feedback loop for educating stakeholders and program personnel on improving “treatment fidelity” and program delivery.

One advantage of the ETMM approach is that the longer time-line (Principle 1) makes it easier to incorporate a formative evaluation phase in the early part of a program’s development, followed

by a summative evaluation at a later stage. Both phases of evaluation should ideally use theory-driven designs (Principle 2).

In textbook descriptions, the results of formative evaluations are purposefully fed back to program personnel and stakeholders to support local decisions and program improvements. Summative evaluations are intended for making more final judgments of a program's worth or merit and support decisions on program continuation/discontinuation, funding, or expansion (Fitzpatrick, Sanders, & Worthen, 2003; Scriven, 1991). Any and all of the studies conducted in the early phases of a program's development (see examples under Principle 1), could be formative in purpose if they help shape the ongoing design, planning, management and delivery of a program.

In ETMM designs, researchers can effectively use the formative evaluation phase to accomplish two objectives. The first objective has to do with enhancing treatment fidelity by promoting evaluation use. The question—Have the results helped move major stakeholders and delivery personnel to the desired actions towards program improvement—must be confronted (Henry & Mark, 2003). Evaluation researchers have a responsibility to be service-oriented; part of this responsibility involves educating stakeholders on how they can modify, tighten, and improve program processes (Joint Committee on Standards for Educational Evaluation, 1994). When accomplished, this objective raises fidelity, consistency, and definition of a “treatment” before experiments are undertaken. At least *one cycle* of formative feedback, thus, may be considered a necessary prerequisite for shaping the field-treatment conditions before summative evaluations are implemented for definitively answering cause-effect questions.

Researchers can also use the formative phase to develop, select and validate effective variable measures—for *all* relevant variables—as a preparation for the summative phase, a second design objective. Poor instrumentation or observation methods can invalidate findings of a summative evaluation on program impact. Stakes are lower in the formative phase, and the opportunity for developing better instrumentation should not be lost (an oft ignored step). Delaying experimental studies thus can help in designing and refining measurement of outcome variables (DVs) as well as observation procedures for treatment and other critical moderators/mediators (IVs).

Principle 4. ETMM designs incorporate sharply focused causal questions in appropriately timed field experiments, and incorporate well-defined treatment and interaction variables in the design.

To be able to answer “What Works” questions with experimental methods, the causal question must incorporate a well-defined, rather than an amorphous treatment condition. Further, Cook (2002) tells us, “at their most elegant, (experiments) can responsibly test only a modest number of interactions between treatments” (Cook, 2002, p. 179). The knowledge gained through the early research on a program (see Principles 1–3) can help researchers identify a small set of sharply formulated causal hypotheses to test from a complex world of primary, secondary, and tertiary causal factors. The experiment should thus be timed and executed more appropriately.

This principle is best illustrated with an example, such as Voucher Plans. It is easy to pose a question like the one that legis-

lators are currently raising: Do vouchers (IV) improve student achievement (DV)? However, the treatment variable “vouchers”, as stated here, is too diffuse in form to be effectively incorporated into an experimental design. It can assume different operational meanings at different levels of policy implementation. At the legislative and highest policy-making levels, it may involve redistribution of dollars to offer parents free choice of schools; at the district and school level, it could shift enrollment numbers, demographic compositions and per pupil expenditures; these factors can, in turn, impact teacher quality and instructional resources in classrooms—all of which could potentially impact student achievement in a school, the DV of interest in the problem.

If one accepts the “program theory” just proposed on vouchers, we have already identified several primary, secondary, and tertiary factors that could interact in unknown ways with other IVs at different sites, generating confounded (or uninterpretable) effects on achievement (the DV) for an unwary researcher. A rush to randomized experiments without understandings of the shape and form that such factors take, would not improve understandings of whether vouchers “work”. Nor would it offer a reasonable test for the effects of the “treatment”, allowing informed inferences about replicability.

Descriptive studies, of the type suggested in the preceding sections (Principles 1–3) permit researchers to not only understand but better define the field-treatment conditions. They help them select the key variables to use as statistical or procedural controls in field settings (those that are the most severe threats to internal validity; Principle 4). They can also help narrow the number of interactions to test among several primary, secondary, and tertiary treatment factors (Principle 4).

Principle 5. ETMM designs effectively combine qualitative and quantitative research evidence to obtain understandings of how, why, and when a program works, and to inform causal interpretations.

Lastly, multiple research methods and tools of inquiry—qualitative, non-experimental, and experimental—are essential arsenal for researchers who attempt studies on “what works” in education. Without effective use of a variety of research methods at appropriate times, the quality of evidence on a program suffers, and interpretations of causality are limited (goes back to Principles 1–4).

For example, in the start-up stages of a new program, a context and program theory analysis may be undertaken by evaluators to investigate the scope and dimensions of the problem that the program was intended to ameliorate (such as, students in a school failing to meet standards on standardized reading tests). Such a study may involve survey research or analysis of existing databases to test initial relational structures; it may also involve qualitative interviews of program managers or delivery personnel (principal and teachers in the school). A study of program implementation may follow (a process evaluation), using direct observation of classroom activities as well as self-report surveys of selected program staff, parents, and students.

Both the above types of studies, while not using experimental methods, would help shed light on a variety of context, process, and input indicators operating at the program site. To empirically test process-outcome links, experimental methods with quantitative outcome measures, would also be necessary. The

timing and design of the experiment will be crucial to informing determinations of causality (Principle 4). Substantive understandings that researchers gain from early phases (Principles 1–3) will provide checks and balances when making causal links. In short, researchers should be prepared to draw from a wide gamut of research methods to understand how well a program “works”.

To do the job well, then, good evaluators would not jump pell-mell into experimental studies, but adopt a long-term evaluation design that recognizes the typical life span of a program during evaluation planning. Rather than singular dependence on experimental methods, evaluation designs would adopt mixed-method approaches. Initial phases of such an evaluation would be geared towards studying context and systemic factors. Early findings would be directed towards shaping and tightening the program at the site(s) with formative evaluations using descriptive and often, qualitative or non-experimental methods (Abma & Stake, 2001; Stake, 1997; Stufflebeam, 2003). Subsequent phases would address program impact issues with more analytic, experimental designs that show deeper understandings of causal contingencies and co-contingencies (Cook, 2002), grounded in actual study of the program’s larger environment and underpinning theory.

Depending on the question to be addressed, a competent evaluation researcher may pick from several research methods, such as ethnography, case study methods, survey research methods, the analysis of archival records, experimental/quasi-experimental methods, historiography, secondary analysis of large databases, or documentary analysis (Yin, 1993, identifies some of these). The choice of methods, however, would be guided by the developmental stage of the program and the most pressing questions to be answered at the time of data collection.

A Case Study Applying the ETMM Design Principles

To illustrate how the five ETMM design principles can be applied to improve the quality of research evidence on a program, an evaluation case is briefly discussed (Chatterji, Kwon, Pacoza, & Sng, 2004). The case represents a small study, limited to one school, 16 classrooms, and roughly 250 students. Because of its smaller scale, it facilitates a closer examination of the merit and usefulness of the ETMM approach. The argument made here, however, is that the ETMM principles would transfer to studies that are large scale as well.

The case focuses on a study of an extended day supplemental program, developed and distributed by a North American curriculum corporation. The study was designed to follow the program for a little over a full academic year (Principle 1), from its time of adoption in the preceding summer term, through the early pilot stage (fall semester, 16 weeks), to a more mature stage of implementation (spring semester, 16 weeks).

The work began with an attempt to analyze the program theory and the program context through a thorough review of materials, videos, documentation supplied by the developers, ongoing consultations with staff of the corporation who were working on site, attendance of parent and teacher orientation/training sessions, and informal interviews with school leaders and teachers (Principle 2). From a research design perspective, the first semester was dedicated to a formative phase with in-depth observations of school and classroom activities, teacher

behaviors, other input/process variables, and preliminary assessments of student outcomes. Early in the second semester, the results of the formative evaluation were presented to participating and non-participating teachers/school staff (Principle 3). The second semester was then used to conduct a summative impact study, taking into account all that was learned in the first phase (Principle 3).

Due to administrative constraints, a true experiment could not be mounted with random assignment of children to treatment and control conditions prior to program implementation. Following the summer orientation and in-service sessions, eight teachers volunteered to participate in the program (their classes became the treatment classrooms); eight other classrooms were matched by grade for the study. Both evaluation phases, formative and summative, thus incorporated quasi-experiments (Principle 4), comparing matched-pairs of children by level (primary or Pre-K through 2, intermediate or grade 4 through 5) and demographic characteristics (gender, socioeconomic status, native language, and ethnicity). To draw conclusions on program effects, the summative phase used the achievement scores from the end-of-Phase 1 as the covariate, and multi-factor ANCOVA analyses with effect size comparisons. In the end, two main factors were examined, level and treatment condition. Several potential moderators and contaminants were empirically evaluated during both phases of the evaluation, and ruled out as possible threats in the final analysis—that is, treatment and control groups were not found to differ on these factors (Principle 4).

Local Benefits of the ETMM Design

Several benefits accrued from applying the ETMM approach. These included the ability to compare qualitative and quantitative evidence to make more balanced and accurate conclusions (Principle 5); to document site-specific departures from the program’s theory, such as interfering classroom management variables; to improve the authenticity of treatment delivery with specific feedback to participant teachers/program staff before initiating the summative study; to empirically test and eliminate extraneous variables, such as the regular curriculum’s alignment with the supplemental program, in planning the statistical design for the summative study; to document changes in the treatment variable over time due to student mobility; to redefine the treatment variable before incorporating it in the ANCOVA; and to develop and validate achievement and survey-based construct measures prior to the summative phase.

A finding of the formative phase had been that in intermediate classrooms, there were zero or negative effects on reading and mathematics achievement outcomes when treatment and comparison group children were compared. Review of the observational data showed that contrary to “program theory”, student misbehavior and associated management issues severely interfered with delivery processes in the intermediate classes during the first semester. The finding led to school-wide discussions once results were shared among stakeholders. In the following semester, the program delivery processes were observed to actually change in grade 4–5 classes. The outcomes also reversed: intermediate students were about six-tenths of a standard deviation unit above comparison students in mathematics, and adjusted for mid-year

performance, they were performing higher than their matched peers (a difference that was statistically significant at the .01 level). Because the qualitative and survey evidence showed that program delivery had been relatively smooth in the second semester, causal conclusions were now more clear and defensible.

After a year-long intervention, the study showed positive effects of the program, but only on tests aligned with the supplemental program's objectives and worksheets. Other significant effects that favored the treatment children were the speed with which they completed worksheet problems in reading and mathematics, and the number of difficult items they attempted. The effects were judged to be "gross effects" (effects of program confounded with teacher and school/developer involvement levels). No positive effects could be documented on other, broader outcome measures, such as the state and city standardized tests. Significant resources had been invested in the first year by the developers and school principal to achieve the observed effects at the site; and the data showed that participating teachers were generally accepting of the program by the end of the year. A recommendation was made to continue the program for longer periods at other sites before the next formal impact evaluation was undertaken.

There were limitations in the ETMM design, as implemented. For instance, teachers by grade were comparable in certification and years of experience, but there was no control for their varying levels of interest or commitment to the program in treatment classes. However, had only a matched groups quasi-experiment been used instead of a two-semester long mixed-methods design (as would occur if the NCLB/WWC's prescriptions were followed), far less would have been understood as to why the program effects were manifested on some outcome measures but not on others, and why findings changed from the formative to the summative phase. Researchers and stakeholders thus had a better grasp of the form in which the program was likely to generate replicable effects at other public school sites. In sum, the ETMM design provided a more comprehensive body of evidence on the workings of the supplemental program.

Recommendations for Revising the WWC's Standards

The current version of WWC's screening instrument, the *Study DIAD*, covers eight main areas in its composite questions: relevance of an intervention, relevance of outcome measures used, fairness of the comparison made, lack of contamination in the research design, generalizability of findings, subgroup effects tested, effect sizes and the statistical precision of the outcome (See w-w-c.org). In the main, all eight areas and specific questions subsumed are presently weighted towards the use of experiments/quasi-experimental methods. While true to positivistic traditions, these criteria fall short in several respects.

A serious rethinking of current criteria for gathering evidence, drawing on more systemic principles of research design and ideas of scientific realism, is recommended here. Additional screening questions, reflecting the five guiding principles of ETMM designs advanced, could include the following.

1. Did researchers use a long-term research plan, attacking a significant period if the life of the program or intervention for study, in test sites?

2. Were relevant environmental and program variables formally documented and studied *in situ* (for comparison, cross-validation, and gaining improved understandings), in the early and later phases of implementation?
3. Were multiple research methods appropriately employed, matched to the developmental stage of the program?
4. Did research methods broaden understandings of the program's theory, and relationships among context, input, process and outcome factors as they operate *in situ*?
5. Did the overall research plan incorporate both a formative and summative phase?
6. Were formative studies conducted with intent to document and shape treatment fidelity *in situ*, before mounting more analytic, experimental methods?
7. Was the use of experimental methods to examine summative program effects informed by findings on all relevant moderators, mediators, and alternate causal agents from the earlier phases of study?
8. Did the design of statistical analyses incorporate relevant causal contingencies and co-contingencies evaluated through prior phases of the evaluation?
9. Was attention paid to establishing valid and reliable measures for all relevant variables (not merely outcomes) prior to summative study of program effects?
10. Were conclusions about causality and program's effects, external validity, and replicability, supported by a holistic appraisal of formally gathered research evidence with multiple methods and over time?

Recently, Eisenhart and Towne (2003) noted that the standards are presently undergoing revisions based on feedback from the public and members of the educational research community (p. 35–36). As of February 2004, regression discontinuity and time series designs were under consideration for inclusion (J. Valentine, personal communication). Despite changes to date, the rather narrow, original structure of the *Study DIAD* remains. Some changes appear to be rather superficial modifications to the language.

Conclusion

The intent of federal policy on scientific evidence is to help summarize the best available research on programs for school-based clients, and thereby improve the quality of education in schools. At a time when schools are being held accountable for high standards and student outcomes, this is a laudable mission. The current criteria, as given in the NCLB legislation and WWC's *Study DIAD*, however fail to capture some of the most important characteristics of well-conducted, policy-relevant evaluation research. If followed, they are likely to have widespread and long-lasting effects on both the quality of educational evaluations, as well as the quality of evidence.

This article forwarded the notion of ETMM designs, drawing from existing theoretical literature and some early thinking of Cronbach and Campbell, with the claim that they will yield *better evidence* than experimental methods employed alone. To close the gap between methodological theory versus practice, it synthesized relevant ideas in the form of five guiding principles for the conduct of ETMM-type studies, offering criteria for evaluating their quality. It argued that ETMM designs permit better conclusions and recommendations on a program's effects. It

pointed out that while the current WWC standards make a good start, they overlook too many well-documented essentials for effective field research.

In this context, the type of study likely to receive future federal and non-government funding, is a significant issue. Despite a widespread mention of ETMM concepts in the professional and theoretical literature (some of it evidenced in academic debates), studies along ETMM lines are *rarely* proposed. Their absence, perhaps, is caused by fear of rejection associated with their heavy resource and time demands. Given the unprecedented federal policy environment impacting educational research today, that course needs to now change.

Large federal grants are already supporting impact studies of widely disparate interventions clustered under the same label—most prematurely employing very large-scale experiments, influenced by the language leaning towards “randomized field trials”. For example, to decide if technology “works”, studies of technology-based instruction have been the focus of large grants. Prior to scaling up to larger multi-site implementation projects, program testing should occur in *small numbers of carefully selected sites*, with tightly conducted ETMM-type designs. If indeed gathering sound evidence on “what works” is the aim, federal government leaders must at least consider ETMM-type designs as a reasonable scientific option.

In conclusion, it is no longer sufficient for the academic and professional research community to continue discussions amongst themselves on methods for obtaining best evidence. It is time to step up, inform, and as necessary, change federal policy, before it is too late. More discussions, thus, should continue on gathering research-based evidence before the WWC’s standards and related federal policy documents are finalized.

NOTES

¹ The IES was previously known as the Office of Education Research and Improvement, OERI.

² I refer here to a statement prepared by the AEA panel represented by Randall Davies, Ernest House, Cheri Levenson, Linda Mabry (chair), Sandra Mathieson and Michael Scriven, and disseminated by the AEA Board and President Richard Krueger in December 2003. The statement was made in response to the U.S. Department of Education’s notice of proposed priority, Federal Register RIN 1890-ZA00, Nov. 4, 2003, on “Scientifically-based evaluation methods”: <http://www.ed.gov/news/fedregister/index.html>.

³ As of this writing, independent research agencies such as SRI International have already posted “Buyers’ Worksheets” on the web targeting school-based audiences such as adopters of instructional technology programs. The development of their latest worksheet (Center of Technology in Learning, SRI International, 2002) was sponsored by the Planning and Evaluation Service, U.S. Department of Education, and mimics the language and criteria for scientific research found in WWC’s DIAD and NCLB, including, for example, use of medical metaphors that do not transfer to educational environments without qualification, such as references to treatment “dosage”.

⁴ Gross effects are effects of a treatment on an outcome variable, confounded with effects of other variables.

REFERENCES

Abma, T. A., & Stake, R. E. (2001). Stake’s responsive evaluation: Core ideas and evolution. *New Directions in Evaluation*, 92, 7–22.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strate-

gic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

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

Bernstein, D. J., Whitsett, M. D., & Mohan, R. (2002). Addressing sponsor and stakeholder needs in the evaluation-authorizing environment. In R. Mohan, et al (Eds.), *New Directions in Evaluation*, 95, 89–99.

Bickman, L. (2000). Summing up program theory. *New Directions in Evaluation*, 87, 103–112.

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

Campbell, D. T., & Stanley, J. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago, IL: Rand McNally.

Chatterji, M. (2002). Models and methods for examining standards-based reforms and accountability initiatives: Have the tools of inquiry answered pressing questions on improving schools? *Review of Educational Research*, 72(3), 345–386.

Chatterji, M., Kwon, Y. A., Paczosa, L., & Sng, C. (2004, April). *Gathering research evidence on a supplemental instruction program: A theory-driven quasi-experiment supported with context/process data*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cook, T. D. (2002). Randomized experiments in education: Why are they so rare? *Educational Evaluation and Policy Analysis*, 24(3), 175–199.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116–126.

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

Cronbach, L. J., & Suppes, P. (Eds.) (1969). *Research for tomorrow’s schools: Disciplined inquiry for education*. A report by the National Academy of Education. Committee on Educational Research. New York: Macmillan.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on “scientifically-based” education research. *Educational Researcher*, 32(7), 31–38.

Erikson, F., & Guitierrez, K. (2002). Culture, rigor, and science in educational research. *Educational Researcher*, 31(8), 21–24.

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

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2003). *Program evaluation: Alternative approaches and practical guidelines* (3rd ed.). White Plains, NY: Longman.

Greene, J. C., & Caracelli, V. J. (1997). (Eds.). Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms. *New Directions in Evaluation*, 74. San Francisco, CA: Jossey Bass Publishers.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation’s influence on attitudes and actions. *American Journal of Evaluation*, 24(3), 293–314.

House, E. R. (1991). Realism in research. *Educational Researcher*, 20(6), 2–9.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression* (3rd ed.). Beverly Hills, CA: Sage Publications.

Joint Committee on Standards for Educational Evaluation, & Sanders, J. R. (1994). *The program evaluation standards: How to assess evaluations of educational programs*. (2nd ed.). Thousand Oaks, CA: Sage Publications.

- Pellegrino, J. W., & Goldman, S. R. (2002). Be careful what you wish for: You may get it: Educational research in the spotlight. *Educational Researcher*, 31(8), 15–17.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach*. (6th ed.). Thousand Oaks, CA: Sage Publications.
- Saloman, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher*, 20(6), 10–18.
- Saxe, L. & Fine, M. (1981). *Social experiments: Methods for design and evaluation*. Beverly Hills, CA: Sage Publications.
- Scriven, M. (1991). *Evaluation thesaurus*. Newbury Park, CA: Sage Publications.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Co.
- Smith, M. F. (1989). *Evaluability assessment: A practical approach*. Norwell, MA: Kluwer Academic Publishers.
- Spector, P. E. (1981). *Research designs*. Beverly Hills, CA: Sage Publications.
- St. Pierre, E. A. (2002). “Science” rejects postmodernism. *Educational Researcher*, 31(8), 25–28.
- Stake, R. E. (1997). Case study methods. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 401–421). Washington, DC: American Educational Research Association.
- Stufflebeam, D. L. (1983). The CIPP model for program evaluation. In G. F. Madaus, M. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation models* (pp. 117–141). Boston, MA: Kluwer-Nijhoff.
- Stufflebeam, D. L. (2003). Evaluation views and roots of Daniel L. Stufflebeam from the perspective of the CIPP model for evaluation. In M. Alkin and C. Christi (Eds.). Unpublished manuscript.
- Suchman, E. A. (1967). *Evaluative research: Principles and practice in public service and social action programs*. New York: Russell Sage Foundation.
- U.S. Congress. (2001). *No Child Left Behind Act of 2001*. Public Law 107-110, 107th Congress. Washington, DC: Government Printing Office.
- Valentine, J. C., & Cooper, H. (2003). *What Works Clearinghouse study design and implementation device* (Version 1.0). Washington, DC: U.S. Department of Education.
- Weiss, C. H. (2000). Which links in which theories shall we evaluate? *New Directions in Evaluation*, 87, 35–45.
- Yin, R. K. (1993). *Applications of case study research*. Newbury Park, CA: Sage Publications.

AUTHOR

MADHABI CHATTERJI is Associate Professor of Measurement, Evaluation and Education at Teachers College, Columbia University, Box 68, 525 W. 120th St., New York, NY 10027; mb1434@columbia.edu. Her areas of specialization are evaluation methods and theory, instrument design and construct validation using classical and Rasch measurement models, and standards-based reforms.

Manuscript submitted February 9, 2004
 Revision received June 23, 2004
 Accepted September 28, 2004
 Reprinted June 2005

Save The Date!

Thursday, October 20, 2005

**Second Annual Brown Lecture
 in Education Research**

Speaker: Claude M. Steele

This Stanford University psychologist
 has changed how social scientists think
 about prejudice and stereotypes.

6 p.m. Lecture, followed by Reception
 Ronald Reagan Building
 and International Trade Center

Washington, D.C.

Details available in September at
www.aera.net