Chapter 22
Speaking Truth to Power with Powerful Results: Impacting Public Awareness and Public Policy

Mack C. Shelley II

This part of the book focuses specifically on the public policy issues of: (a) the ways in which global education funding patterns reflect governmental—and perhaps societal—priorities; (b) the role of research ethics boards in enforcing public policy norms regarding what is appropriate for science and literacy education research; (c) rules and expectations established by national legislative action and by professional associations for maintaining the security of the voluminous sets of data needed for sustained research excellence in science and literacy education research; (d) how qualitative research studies can be employed to provide broader and more lasting impacts on public policy making through systematic research reviews, secondary analysis, comparative case studies, and metasynthesis; and (e) how Gold Standard(s) inform education experts and policy makers about what should be done with research findings. This chapter is intended to elaborate many of the points that have been made earlier in this book and perhaps to foreshadow an action agenda for education researchers and those who seek to influence the shape and direction of public policy. One of the major lines of argument is the need for eclecticism—in methodology, subject matter expertise, and policy agendas. Consistent with that theme of the virtue and necessity of eclectic approaches, and to honor the need for truth in advertising, it may be helpful to know that the author of this chapter is a faculty member with a joint appointment in a department of statistics and in a department of political science, with about 30 years of experience with statistical consulting, and with a background in public policy, program evaluation, and public administration. That background may help explain where this chapter is coming from—as a somewhat eclectic, multifaceted exploration of a topic that is very much at the interface of several disciplines and multiple research methodologies.

M.C. Shelley II
Iowa State University

M.C. Shelley II et al. (eds.), *Quality Research in Literacy and Science Education*,

22.1 Speaking Truth to Power

One of the central points of the study of public policy, political science, and public administration is the artistry required to *speak truth to power*. As expressed in particular by Wildavsky (1979), the process by which experts convey the gravamen of their findings to the *powers that be* who make and enforce decisions that may be driven by those research results is an art and craft that reinforces the science and practice of politics. In highly abbreviated form, the essential point is how to reach across the gulf that is created by an unequal distribution of power (researchers having rather little and decision makers having very much more) to transmit understanding to those who are able to compel binding decisions. This involves, among other traits, the refusal to be intimidated by the presence of power, the commitment to pass on knowledge even to an audience that may not be appreciative, and the willingness and artistry to explain inconvenient truths to those who may be shown to be wrong—for example, Mathematica Policy Research showing the possible ineffectiveness of sexual abstinence education programs compared to traditional sex education programs (Trenholm et al., 2008) or the Institute of Education Sciences (IES) concluding that the federally funded and officially endorsed Reading First initiative was no better for student outcomes than alternative literacy programs (Gamse, Bloom, Kemple, & Jacob, 2008).

An important aspect of the increasingly sophisticated evaluation efforts required of scientifically based research standards is the need to strike a balance between stakeholders (such as school district administrators) and the accountability systems that require specialized expertise and that can complicate the process of speaking truth to power (Schmitt & Whitsett, 2008). Cohn (2006) noted that the tradition of scholarly detachment has led to the perception that it is difficult for academics to implement the ideas and advice they have afforded to the policy-making powers that be. He argued that academics can and should make more effective use of the opportunities that are available to them to influence public policy and that policy makers can make better use of scholarly expertise through third-community, public- and private-sector actors who influence or advise policy makers by producing and disseminating usable policy alternatives. These policy advisers include members of the research staffs of government ministries, cabinet committees, central agencies, task forces, investigatory commissions, public inquiries, research councils, private consulting organizations, political parties, interest groups, and think tanks. Cohn emphasized that academics must be sensitive to the need to join in the efforts of advocacy coalitions to situate policy decisions at the political moment when sufficient support exists for a decision to be made.

In the genre of political science, Kingdon (1995) developed a thorough conceptualization of what it takes for an idea whose time has come to make it to the decision-making phase of the policy process. Kingdon's framework uses the metaphors of the policy primeval soup and the confluence of three streams—a political stream related to elections, pressure group actions, and swings in public opinion; a policy stream, in which a policy proposal emerges as the best available alternative; and a problem stream, in which a problem emerges that is seen as important—feeding into the making of public policy by getting an issue onto the policy agenda. Kingdon's perspective emphasizes the essentiality of getting on the policy agenda by making sure a problem and its possible solutions become identified as an issue that requires public-sector attention, discussion, and action. Certainly, education issues generally are high visibility and frequently are caught up in the flow of the currents and cross-currents streaming into, through, and from the policy process. Navigating successfully the shoals and eddies of these streams, and the occasional Odyssean adventures through Scylla and Charybdis, is not for the faint of heart and requires more than the usual degree of commitment to persevere through to success.

Henig (2008) argued that, together with the old image of the ivory-tower aloof academic, "the old model of 'speaking truth to power' in which the scholar as favored advisor whispers into the ear of elite leaders, also is passé; in the age of mass media and the Internet, discourse about research has been democratized" (p. 360). This certainly does seem to be a contemporary assessment of the current state of speaking truth to power; but, far from negating the basic premise of the Wildavsky argument, it modernizes an already well-established perspective on politics, society, and how research interfaces with realities as perceived both within the corridors of power and by the public. Henig surely is correct in noting the need for academic "buffers against ideology and the politicization of the knowledge enterprise [to help maintain] a distinction between research and advocacy, between pursuit of knowledge and pursuit of advantage, between sounding good and being right" (p. 360).

Widespread dissemination and accurate interpretation of the results of education research also depends on contemporary media outlets being staffed by reporters who have sufficient background to know quality results when they see them and who are able to focus on the importance of the findings over the more headline-grabbing controversies that all too often are the natural target of media efforts to reflect or influence policy makers' opinions (Rotherham, 2008). Furthermore, academics need to be aware of the basic constraints, practices, and genre of popular media: 10-second sound bites, brief video clips, and journalistic versions of research reports of interspersed claims, evidence, and narrative that all too frequently imply applications and a degree of certainty that may not have been intended by the original researchers.

The utility of research results certainly needs to be enhanced. Brewer and Goldhaber (2008) argued that:

> since most consumers of the work will not have the time or capacity to judge its quality … [for] the rigor and relevance of educational research … to be increased, we will need a concerted effort from both consumers of research and suppliers who recognize the desperate need for improvement. (p. 364)

Getting the attention of education leaders and convincing them to make productive use of research results surely is enhanced when the research results are consistent, demonstrably relevant to the needs of educational practitioners, and disseminated quickly. That process is facilitated when fostering data literacy is a priority of school leadership and when consensus emerges on the appropriate research design strategy (Fusarelli, 2008; see also Ingersoll, 2008, on out-of-field teaching; and Kim, 2008, on reading research). Kim concluded optimistically that:

> we will be able to establish norms of excellent practice rooted in scientific research and governed by a community of peers. Ultimately, teachers must have access to truth and power if they are to create professional norms that nurture effective instruction and support efforts to help children become proficient readers. (p. 375)

Throughout this book, and perhaps especially in the chapters that constitute Part IV, the authors have addressed a multiplicity of the facets at the interface between power and expertise—where public policy joins with expert judgment and academic expertise to synergize the politics of knowledge (Hess, 2008). Hess argued that, in contrast to health care research, the record of education research is less replete with success stories, and hence "educational research has not earned similar trust or good will, and its advocates have been unsuccessful in making the case that research ought to be funded despite its painstaking pace and uncertain fruits" (p. 356). Henig (2008), going further, noted that "[a]mong policy makers and many scholars, educational research has a reputation of being amateurish, unscientific, and generally beside the point" (p. 357) and thus has less impact than it should, particularly given the internecine methodological disputations that further dispel the idea that education researchers really know what they are doing and that they know how to make proper meaning of the results.

The realization that politics plays a role in the process by which research is filtered and possibly impacts decision making certainly does not surprise the average, randomly selected, social scientist, particularly anyone who may be a card-carrying political scientist. The dimensions of this policy–politics nexus, however, may not be so thoroughly familiar to education researchers or to others who do not reflexively tune in to C-SPAN or other media-generated sources of eye-glaze to those less afflicted with the *can't-help-it* impulse to see and listen to the political process that Iron Chancellor Otto von Bismarck famously likened to sausage-making. Henig (2008) noted that the pressure to produce timely results to fit the dictates of political decision-making schedules:

> is especially the case in politically charged arenas in which groups with tactical interests in advancing or blocking specific policy actions can co-opt the process. Researchers may acknowledge the limitations of their own data and design, but those caveats are often the first things to be stripped from the message as others take it up. In practice, research that aligns with ideological cleavages is more likely to be pushed into the public realm, thus blurring the distinction between advocacy and unbiased analysis. (p. 358)

The final report of the National Mathematics Advisory Panel (NMAC, 2008), based in part on the assessment of 16,000 research publications, provided a recent example of how federal education policy can be impacted by expert panel recommendations. Convened by US President George W. Bush, the panel was formed to advise the administration on how to enhance mathematics education, with members including prestigious professors of mathematics and psychology, a middle school teacher of mathematics, and the president of the National Council of Teachers of

Mathematics. The report concluded, in part, that long-festering debates about what curricular policy to recommend are largely irrelevant:

> To prepare students for Algebra, the curriculum must simultaneously develop conceptual understanding, computational fluency, and problem-solving skills. Debates regarding the relative importance of these aspects of mathematical knowledge are misguided. These capabilities are mutually supportive, each facilitating learning of the others. Teachers should emphasize these interrelations; taken together, conceptual understanding of mathematical operations, fluent execution of procedures, and fast access to number combinations jointly support effective and efficient problem solving. (p. xix)

Similarly, the report found that intense and long-standing policy debates about the relative superiority of teacher-directed or student-directed mathematics instruction miss the point and concluded that:

> [i]nstructional practice should be informed by high-quality research, when available, and by the best professional judgment and experience of accomplished classroom teachers. High-quality research does not support the contention that instruction should be either entirely 'student-centered' or 'teacher-directed'. Research indicates that some forms of particular instructional practices can have a positive impact under specified conditions. (p. 11)

Clearly, high-quality evidence is essential, but not sufficient, in making and justifying instructional decisions. Knowing what to believe, and therefore having a better idea of what to do, is an essential prerequisite for wise public policy making. Synthesizing results across multiple, and often contradictory, studies is a form of high art requiring tools and perspectives that are not readily understandable to many researchers, let alone those who make education policy. To determine which education programs work and, therefore, deserve continued or enhanced support, Slavin (2008a) suggested the following criteria essential for valid program evaluation research: "Clear, thoughtful syntheses in many areas are crucial to providing practitioners, policy makers, and researchers with valid information they can use with confidence to address the real problems of educating all children" (p. 13). As evidenced by recent debates within the education research literature (e.g., Slavin, 2008a, 2008b, and others discussed below), several major efforts to synthesize the current state-of-the-art research record provide the foundation for intentional overviews of research results, including:

- What Works Clearinghouse (US IES, n.d.-b), officially supported by the IES of the US Department of Education (US ED) and now managed by Mathematica Policy Research, Inc.
- Best Evidence Encyclopedia (BEE, n.d.), a collaboration between the Center for Data-Driven Reform in Education in the US ED and Johns Hopkins University.
- Comprehensive School Reform Quality Center (CSRQ, 2006), active from 2003–2006 through the American Institutes for Research.
- The international Campbell Collaboration (Campbell Collaboration, n.d.).
- The United Kingdom's government-supported Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre, n.d.).

These organizations can provide collective results supported by evidence from a broad array of studies, but unfortunately evidence alone does not lead directly to policy decisions. Evidence must interact and react with the decision makers' beliefs, values, and priorities to result in evidence-based policies (see Phillips, Norris, & Macnab, Chap. 27).

22.2 A Theme to Consider: Challenge and Response

Borrowing very loosely from Toynbee's (1934–1961) 12-volume exposition of what he conceptualized as the challenge-and-response cycle throughout recorded human history, the next sections of this chapter lay out what may be regarded as the challenge posed by the present state of affairs of Gold Standard(s) expectations for education research and the response that has come, and that may be expected (or hoped for), from that research community. In doing so, the intention is to provide a broad context within which to consider the implications for education research and public policy agendas.

22.2.1 The Challenge

In the United States, and in many other countries, research funding from government agencies and other sources increasingly has become tied more closely to use of the medical model of randomized clinical trials (RCTs), featuring: (a) randomized assignment of individual subjects or clusters of subjects to treatment or control groups, (b) the need to ensure fidelity of treatment effects over both space and time, and (c) consistent and accurate measurement of well-defined outcomes. This focus on RCT-style interventions recently has been emphasized in the requirements for research in education and in other human sciences. The emphasis on the expressed needs for randomization, control, and measurement has led to a greater need for careful attention to the requirements of focused research by content experts in many diverse aspects of education inquiry and for research methods experts to be willing and available to partner in joint efforts with content specialists. These partnerships are not always easy or straightforward—particularly when there is not a lot of overlap in the substantive knowledge base and the methodological expertise of those participating in these joint ventures.

Knowing how to apply the logic of experimental and quasi-experimental methods has become essential for the successful pursuit of research awards from government sources—in the United States, from public agencies such as the US ED, the National Institutes of Health (NIH), and the National Science Foundation (NSF)—and many other funding sources (e.g., W. T. Grant Foundation, Spencer Foundation). Furthermore, it is important for successful publication of the results from such studies in appropriately high-level outlets. Competition for funding from these and other sources generally has become much fiercer; for example, what once was about a one-in-three reasonable prospect of succeeding with a grant proposal submitted to NSF now is more like a one-in-ten shot in the dark.

22.2.2 The Response

So, how has the education research community begun to address this challenge? The need to deal with the current and future situation has become a major point of discussion among researchers who are content experts in education and the social sciences, often in conjunction with their qualitative and quantitative research methodology colleagues. One recent example is the Ragin, Nagel, and White (2004) NSF-funded volume based on a workshop on the scientific foundations of qualitative research. This publication provides essential recommendations to improve the quality of qualitative research proposals and for evaluating the scientific and substantive merits of such proposals. Among the key questions addressed is: what is an ideal qualitative proposal? Ragin and colleagues also recommended how NSF (and implicitly any other public agency) can support and strengthen high-quality qualitative research, especially in light of the specific resource needs of qualitative researchers that may be understood less well by the reviewers of qualitative proposals than reviewers understand the research needs for more traditional, quantitative submissions.

The Ragin and colleagues' (2004) document provided a substantial set of recommendations for designing and evaluating an ideal qualitative research proposal "to improve the quality of qualitative research proposals and to provide reviewers with some specific criteria for evaluating proposals for qualitative research" (p. 3). It is understood that not all of these challenging targets can be met and that what is sauce for the qualitative goose also is sauce for the quantitative gander, which can be applied to research proposals of all methodological persuasions. The recommendations include the following:

- Write clearly and engagingly for a broad audience
- Situate the research in relation to existing theory
- Locate the research in the relevant literature
- Articulate the potential theoretical contribution of the research
- Outline clearly the research procedures
- Provide evidence of the project's feasibility
- Provide a description of the data to be collected
- Discuss the plan for data analysis
- Describe a strategy to refine the concepts and construct theory
- Include plans to look for and interpret disconfirming evidence
- Assess the possible impact of the researcher's presence & biography
- Provide information about research replicability
- Describe the plan to archive the data (pp. 3–4)

A second set of recommendations addresses ways in which the research grants process can better support and strengthen qualitative research and enhance the productivity of qualitative researchers, taking into account particularly the resource needs of qualitative researchers:

- Solicit proposals for workshops and research groups on cutting-edge topics in qualitative research methods

- Encourage investigators to propose qualitative methods training [professional development and education]
- Provide funding for such opportunities to improve qualitative research training [professional development and education]
- Inform potential investigators, reviewers, and panelists of qualitative proposal review criteria
- Give consideration, contingent upon particular projects, to fund release time for qualitative researchers beyond the traditional 2 summer months
- Fund long-term research projects beyond the traditional 24 months
- Continue to support qualitative dissertation research
- Continue to support fieldwork in multiple sites (Ragin et al., 2004, p. 4)

On the quantitative side of the research methodology spectrum, the American Statistical Association (ASA, 2007), in collaboration with NSF, produced the seminal report *Using Statistics Effectively in Mathematics Education Research*. This publication was the product of 3 years of NSF-funded workshops conducted by the ASA's Working Group on Statistics in Mathematics Education Research, whose membership included leading experts in mathematics education, psychology, measurement, and statistics. Five steps to effective quantitative research in education were noted, which are summarized below.

Step 1: Generate research ideas. Recommendations to researchers include:

- Identify ideas and questions about a topic of interest.
- Determine specific research questions to investigate.
- Build an argument about why this question is worth investigating.
- Make the researchers' beliefs and assumptions about the topic explicit.
- Examine primary and secondary research literature to clarify the researchers' beliefs, biases, and assumptions about the research topic.
- Review existing research and nonresearch literature to determine the current state of knowledge about the questions.
- Determine the concepts and constructs associated with the topic; develop a conceptual framework linking the concepts and constructs.
- Identify research methods (e.g., experimental methods, cognitive models, participant observation) that can provide information about the concepts and constructs.
- Synthesize knowledge about the research question to date.

Step 2: Frame the research program, considering goals and constructs, measurement, and logistics and feasibility.

(a) Goals and constructs recommendations include:

  - Propose a conceptual model linking the constructs.
  - Explore existing data and observations.
  - Identify relevant variables and define them operationally.
  - Use past data and observations to develop potential hypotheses.
  - Determine appropriate research methods.
  - Identify relevant measures or the need for new measures.

- Gather exploratory empirical data to test the research framework.
- Formulate a research question; outline a plan to answer the question.
- Discuss the possibility that measures (e.g., gain scores) may lead to faulty conclusions.
- Provide exploratory and descriptive statistics with appropriate graphs and interpretations.

(b) Measurement recommendations focused on developing and reporting on assessment measures used in education research that have the qualities of validity (the extent to which a measure is meaningful, relevant, and useful for the research at hand), reliability (the extent to which the measure is free of random error), and fairness (the extent to which measures are implemented consistently and validly for all subgroups) include:

- Examine previously used measures; decide if it is necessary to create new ones.
- Provide key details regarding development of new measures and/or selection of off-the-shelf measures.
- Report the relationships each variable has with other variables used in the research.
- Explain how measures align with the goals of the research.
- Determine the sample or population from which measures will be obtained.

(c) Logistics and feasibility recommendations include:

- Consider potential ethical issues and risks associated with the proposed interventions.
- Document and test the procedures to be used in an intervention study.
- Design and conduct a qualitative component to assess measurement difficulties and possible lack of feasibility of the study.
- Investigate how to deal with problems, such as study dropouts and missing data.
- Examine and evaluate threats to internal and external validity.
- Develop trust within the research setting.
- Search for useful common measures that can be related to other research.
- Develop, if necessary, tests to determine interrater reliability and internal validity; refine measures.
- Pilot all instruments in an informal setting; conduct a formal field test or pilot study.
- Develop a plan for the formative evaluation of an intervention.
- Meet institutional review board guidelines, ensuring confidentiality and informed consent.
- Anticipate problems in the field; develop an affordable contingency plan.
- Develop a work plan to coordinate measurement and evaluation within an individual site or among multiple sites.
- Determine any demographic differences between the population and the sample studied.
- Describe the method of sampling, if any.

- Identify the sampling unit and the unit of analysis.
- Describe the treatment and measures in enough detail to allow replication.
- Make sure that adequate time, training, and support services exist to perform the study.

Step 3: Examine the research program. By establishing efficacy, the research program can progress to studies that may be able to establish causal patterns. Recommendations include:

- Specify a study design and the associated data analysis plan.
- Identify subpopulations of interest.
- Define the setting in which the study is to be conducted.
- Identify sources of (extraneous) variability; take steps to control variability.
- Refine measures based on research experience.
- Assess the potential portability of measures to broader contexts.
- Ensure that the intervention received by one subject is independent of the person administering it and independent of the other intervention recipients.
- Provide estimates of statistical parameters as well as the results of hypothesis testing.

Special care must be taken to ensure that statistical results are understandable, correct, and interpreted appropriately. For formal statistical inference, researchers should:

- State the hypotheses clearly.
- Specify a statistical model that addresses the research question.
- Define the population of interest and exclusion/inclusion criteria.
- Describe the characteristics of the study sample.
- Describe how random assignment or random selection was used.
- Describe whether implementation was carried out appropriately.
- Explain measures taken to minimize bias.
- Report statistical power and effect size results.
- Report response rates.
- Provide margins of error or confidence intervals.
- Explain how missing data were handled.
- Describe adjustments to minimize the risk of false positive results from multiple tests.
- Summarize the results of tests of assumptions and diagnostic (e.g., goodness of fit) tests.
- Provide sufficient information to replicate the analysis.
- Consider how to link with other databases.

Step 4: Generalize the research program. This usually involves ramping up to larger studies that randomize classes, groups, or individual subjects to the intervention with appropriate within-study controls on the measurement processes to allow the strongest possible interpretation of causal relationships. Recommendations include:

- Assess the potential portability of measures to multiple institutions in a wide variety of social contexts.

- Design and conduct a multi-institutional randomized study.
- Design and conduct a quasi-experiment.
- Conduct a rigorous statistical analysis of the quantitative results of a multi-institutional study (e.g., a survey, an experiment, an observational study) using statistical methods appropriate to the unit of analysis.
- Specify outcomes: intermediate outcomes (goals) and primary and secondary outcomes.
- Specify how covariates were defined, measured, and used.
- Detail appropriate research designs to test the hypothesis (e.g., experiment, quasi-experiment, matching, repeated measures).

Step 5: Extend the research program. A rigorous, generalized study can be achieved by, for example, syntheses of multiple studies, longitudinal studies of long-term effects, and developing policies for implementation. Ongoing formative evaluations are essential to inform the research team about necessary research adjustments and how to improve measures and procedures. Recommendations include:

- Design and conduct a longitudinal study that allows rigorous statistical inferences over time and long-term improvements in curriculum and student performance.
- Describe the nature of the long-term study (e.g., experimental, quasi-experimental, sample survey, observational).
- Describe the rate of dropouts over time and how this was handled in the analysis.
- Describe how the study maintained measurement integrity over time and in different circumstances.

The linkages connecting methodological research sophistication—whether of qualitative, quantitative, or mixed lineage—with content research expertise and the public policy implications of the results of that research have been drawn out by authors representing a broad range of disciplines and sharing a commitment to ensuring that elected and appointed powers can understand the import of the research and make appropriate use of those findings in formulating public policy decisions. As a case in point, the January/February 2008 issue of the American Educational Research Association's *Educational Researcher* (the contents of which are cited extensively below) offers a full spread of articles revolving around measurement issues that arise in the often tricky business of synthesizing the results of multiple educational program evaluations. This issue features a lead article by Robert Slavin (2008a), with replies by Derek Briggs (2008), Madhabi Chatterji (2008), Mark Dynarski (2008), Judith Green and Audra Skukauskaité (2008), and Finbarr Sloane (2008), with a response by Slavin (2008b).

Slavin's (2008a) argument is that syntheses of research on educational programs have become more important for affecting public policy. Thus, it is increasingly important for such syntheses to produce reliable, unbiased, and meaningful evidence-based interpretations of program results. The number of evaluations of any given program tends to be small, so it is essential to minimize bias in reviewing each study. This is achieved by exercising great care in determining and explaining research design, sample size, any adjustments that may have been made for pretest differences, how long the study lasted, effect sizes, and the number of relevant studies.

Careful research synthesis can result in more meaningful ratings of the strength of evidence for the effectiveness of each program. Particularly for researchers who invest heavily in comparing results across multiple studies and make use of various forms of meta-analysis, this is a must-read opportunity.

Another example of the discussion/debate regarding the role of education research in impacting public policy is afforded in the January 2008 edition of *Phi Delta Kappan* (PDK) on the "Politics of Knowledge," which addresses how educational research may be used to inform policy decisions and foster democratic government. This issue manifests various views—many of which are explored in other sections of this chapter and book—of the education policy process from a number of disciplinary perspectives, including political science, economics, policy studies, urban studies, public affairs, educational leadership and policy studies, sociology, and wonkish think tanks. The present book offers its own contribution to the growing volume of literature on education and policy research provided by these and many other authors and outlets.

22.3 What's all the fuss about, anyway? A Brief Backgrounder

In policy circles, it is pretty much de rigueur to provide a background summary of why we are all gathered together to address any given policy issue. Here is a quick overview, as well as a reminder, of essential points that arise in the debate surrounding Gold Standard(s), building on comments made earlier in this book.

Standards for acceptable and, particularly, fundable research, especially in the context of the US ED, have been affected greatly by two major policy innovations: the No Child Left Behind Act of 2001 (NCLB, 2002) passed on January 8, 2002, and the Education Sciences Reform Act of 2002 (ESRA, 2002) passed on January 23, 2002. The latter of these statutes resulted in creation of the IES (US IES, n.d.-a) in the US ED. Together, these developments have reconstituted federal support for research and dissemination of information in education with ramifications for education research in other countries; they are meant to foster scientifically valid research and have established what often is referred to as the Gold Standard for research in education.

These and other developments denote that greater emphasis in fundable education research now is placed on quantification, the use of randomized trials, and the selection of valid control groups. To meet this challenge, there is an obvious need for experts in research design and research methods to work together with content experts, to apply appropriate methods of measurement, analysis, and interpretation.

NCLB was identified in the legislation as "An Act to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind" (NCLB, 2002, para. 1); hence, the eponymous label of the law, which was officially the 2002 reauthorization of the Elementary and Secondary Education Act of 1965. The NCLB established standards for academic assessments in mathematics, reading or language arts, and science; it required multiple, up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding.

ESRA, or HR 3801, defined *scientifically based research* and *scientifically valid education evaluation* standards to apply rigorous, systematic, and objective methodology to obtain reliable and valid knowledge relevant to education activities and programs, and to present findings and make claims that are appropriate to and supported by the methods that have been employed. In HR 3801, scientifically based research includes systematic, empirical methods that draw on observation or experiment; data analyses that are adequate to support the general findings; measurements or observational methods that provide reliable data; making claims of causal relationships only in random assignment experiments or other designs that substantially eliminate plausible competing explanations; replication or the opportunity to build systematically on the findings of the research; obtaining acceptance by a rigorous, objective, and scientific review; and research designs and methods appropriate to the research question posed.

HR 3801 also specified that scientifically valid education evaluation adheres to the highest possible standards of quality with respect to research design and statistical analysis; provides an adequate description of the programs evaluated and, to the extent possible, examines the relationship between program implementation and program impacts; provides an analysis of the results achieved by the program with respect to its projected effects; employs experimental designs using random assignment when feasible and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible; and may study program implementation through a combination of scientifically valid and reliable methods.

Other countries have attempted to enhance educational research using other quality assurance approaches (see Coll et al., Chap. 6). Their approaches are not driven by prescriptive government policy regarding appropriate research approaches, but they do assess the quality of sponsoring institutions, researchers, and proposals in a variety of ways. Funding agencies use evaluation criteria as another tool to facilitate or restrict research approaches (see She et al., Chap. 23).

22.4 How Does the Gold Standard Connect with Public Policy?

Much of the policy debate swirling around implementation of NCLB and the overall US federal government effort to upgrade the quality of education research is related to the creation of IES as the research arm of the US ED through HR 3801. Its mission is to expand knowledge and provide information on the condition of education, practices that improve academic achievement, and the effectiveness of federal and other education programs. Its expressed goal is the transformation of education into an evidence-based field in which decision makers routinely seek out the best available research and data before adopting programs or practices that will affect significant numbers of students (see Hayward & Phillips, Chap. 7).

Perhaps the best articulation of what is meant by the concept and implementation of the Gold Standard for research is provided by the IES's National Center for

Education Evaluation and Regional Assistance (US ED, 2003) user-friendly guide to identifying and implementing educational practices supported by rigorous evidence. The rules of evidence of education interventions come in two levels. The quality of studies needed to establish *strong* evidence requires: (a) RCTs that are well-designed and implemented, following a medical clinical trials model; and (b) that the quantity of evidence needed spans trials showing effectiveness in two or more typical school settings. *Possible* evidence may include: (a) RCTs whose quality/quantity are good but fall short of strong evidence; (b) and/or comparison-group studies in which the intervention and comparison groups are very closely matched in academic achievement, demographics, and other characteristics.

Evaluating whether an intervention is backed by strong evidence of effectiveness hinges on well-designed and well-implemented RCTs, demonstrating that there are no systematic differences between intervention and control groups before the intervention, using measures and instruments of proven validity, and demonstrating the presence of real-world objective measures of the outcomes the intervention is designed to affect. The benchmarks for evaluating whether an intervention is backed by strong evidence of effectiveness include attrition of no more than 25% of the original sample in longitudinal studies, effect size measures of the estimated amount of impact, and p values at the traditional level of 0.05 or less, adequate sample size to achieve statistical significance, and controlled trials implemented in more than one site representing a cross section of all schools.

For researchers in search of guidance on the essential quantifiable aspects of research design, an excellent source is the W. T. Grant Foundation's Optimal Design software (Raudenbush, Spybrook, Liu, Congdon, & Martinez, 2006; W. T. Grant Foundation and University of Michigan, n.d.). Excellent guidance also is provided by Lenth's (2006) Java applets for determining statistical power (Murphy & Myors, 2003; Schochet, 2005) and sample size. Additional guidance is available online through the What Works Clearinghouse (WWC, US IES, n.d.-b) and Campbell Collaboration (Campbell Collaboration, n.d.).

WWC was established in 2002 by IES to provide educators, policy makers, and the public with a central and trusted source of scientific evidence of what works in education. It reviews and reports on existing studies of interventions (education programs, products, practices, and policies) in selected topic areas that apply standards that follow scientifically valid criteria for determining the effectiveness of these interventions. It also provides technical assistance and a registry of evaluators (US IES, n.d.-c) as well as technical working papers (US IES, n.d.-d). These online assessments and documentation are reviewed by a Technical Advisory Group. As of this writing, WWC has provided detailed results for programs in (a) beginning reading, (b) early childhood education, (c) elementary school mathematics, (d) middle school mathematics, (e) character education, (f) dropout prevention, and (g) English language learning. The most fully elaborated information is available on the first four topics. In each area, WWC evaluates program effectiveness as: meets evidence standards, meets evidence standards with reservations, or does not meet evidence screens. Each specific intervention program is evaluated as having: positive effects, potentially positive effects, mixed effects, no discernible effects, potentially negative effects, or negative effects.

22.5 A Possible Template for Science and Literacy Education Research?

So, in this climate what guidance can be provided to education researchers, particularly those in the fields of science education and literacy research? One possible template for how best to impact the science and literacy education policy areas may be afforded in the aforementioned ASA report (2007). Although focused on mathematics education, it offers some suggestions for research and guidance for actions that may be particularly helpful in the process of trying to speak truth to power.

The focus of the mathematics education template is on how best to cumulate the results of a larger corpus of individual studies to achieve a potentially high-impact summary of programmatic interventions in education. This involves consistent and appropriate use of interventions, observation and measurement tools, data collection techniques, and data analysis methods, and consistent reporting of research results. Doing so facilitates replication (or at least another look at the same problem) and, therefore, makes it more feasible to progress toward the goal of achieving a cumulative discipline, which is commonly seen as a hallmark of *science*. To achieve this goal requires both methodological rigor and methodological diversity, as elaborated, for example, by Raudenbush (2005) and the US National Research Council (US NRC, 2002).

The concept of using larger bodies of studies to help inform policy makers through wider application of both quantitative and qualitative forms of integration and synthesis requires further elaboration. In particular, it is essential to highlight the difference between meta-analysis based on aggregating clinical trials conducted for medical research and meta-analysis conducted using the often more tenuous results of education research. Research protocols, and in many cases measurement procedures and data analysis methods, often are well established and—although often couched in highly technical terminology—frequently are understandable to the general public and are explained and interpreted routinely by mass media outlets. In addition, medical experts are frequent visitors to US government executive-branch agencies and to congressional committees; medical experts generally are given a positive reception, indicating that their expertise is widely understood and respected—if not unchallenged.

In contrast to the apparent near-certainty of the results of medical trials, particularly when aggregated across relatively large numbers of broadly similar studies, the reception often afforded to educational researchers frequently is much less positive. This divergence in the amount of slack given to education—as opposed to medical—researchers by the mass media, the general public, and policy makers may be attributable in large part to the diversity and variety of ideological and methodological positions adopted by educational researchers, exacerbated by disciplinary differences among those who conduct research in education spanning higher education, preK-12, sociology, political science, economics, psychology, statistics, and other areas of expertise that may not speak the same language or use the same procedures or methodologies. Another consequential difference is that most medical clinical trials research is designed to measure the impact of drugs administered in usually carefully controlled environments, such that the analysis conducted on the data resulting from those studies often is not complicated by the need to control statistically for other, potentially confounding, variables. In contrast, even with randomized cluster trials conducted at multiple sites, the analysis of data from education research often needs to be adjusted with covariates and frequently is based on outcome measures that are less precise than what can be achieved under clinical trials laboratory conditions. The greater difficulty in achieving sustained precision of measurement and clear data analysis is compounded further by the lesser amounts of funding available for many education experiments or quasi-experiments; if a standard, medical-style, clinical trials experiment is funded at something like $15 million spread over 5 years, the typical funding for an education intervention study is likely to be much less in total amount and may not be sustained for as long.

Building on the themes of methodological diversity and the need for cumulative findings to maximize the impact of those findings on education-relevant public policy, well-established methods exist in the literature on meta-analysis (e.g., Cooper & Hedges, 1994; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001) and multilevel models (Arnold, 1992; Bock, 1989; Bryk & Raudenbush, 1987, 1989, 1992; Goldstein, 2003; Hedges, 2007a, 2007b; Lee & Bryk, 1989; Raudenbush & Bryk, 1986, 2002) accomplish these goals, by aggregating quantitative results across contexts and across units of measurement (such as individual students, classrooms, or districts). A particularly fruitful line of research is to adjust for aggregate setting effects on student outcomes in cluster randomized trials (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000), whereby schools or districts constitute the units that are randomized and individual student results are aggregated and compared across those settings of intervention (e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2007; Raudenbush, 1997; Raudenbush & Liu, 2001; Raudenbush, Martinez, & Spybrook, 2007).

Borrowing from the ASA (2007), if research in science education and literacy is to have more effective influence on policy and practice, it must become more cumulative in nature, as suggested for mathematics education. This requires building on existing research to produce a more coherent body of work. Education researchers must be free to pursue problems and questions that are of interest to them. To influence practice, however, the work must be situated within a larger corpus. There is power in numbers—both in the number of studies and in the number of researchers agreeing with each other. Cumulating studies through consistent use of interventions, observation and measurement tools, data collection techniques, data analysis methods, and reporting of research results facilitates replication (or at least another look at the same problem).

Based on the results of pilot studies and the use of appropriate methods for data collection and analysis, the goal is to generalize the findings from a research program. This is accomplished by first establishing the efficacy of the study then determining its portability (Will it work the same anywhere?) and scalability (Will it work the same when we do this big time?). Extending the research program is best accomplished by synthesizing multiple studies through methods of meta-analysis, conducting longitudinal studies of long-term effects using growth-curve models, and developing an implementation policy that can get large-scale funding and political support. Ongoing formative evaluations are needed, to permit mid-course corrections if they are needed. This is the payoff of speaking truth to power successfully.

Research methods expertise comes in extremely handy in this process. As something of a shameless advertisement, it often is a very good idea to add a research methods specialist to a team writing grant proposals. Doing so often helps improve the prospects for obtaining funding. It usually helps with establishing the rigor of the research design and may help get results that may be listened to by the powers that be. The methodological bag of tricks includes, for example, expertise in focus groups, document and content analysis, interview strategies, logic models, experimental and quasi-experimental design, and working with large, complex, and/or messy databases with methods such as data mining (see Wang, Dziuban, Cook, & Moskal, Chap. 19; Ye, 2003). In particular, quantitative methodologists know how to handle the nearly inevitable complications that arise from the presence of missing data through the use of imputation, plausible values, survey weights, poststratification, and other mechanisms.

Complex contemporary methods of data analysis that convey powerful results to policy makers include hierarchical linear modeling (e.g., Arnold, 1992; Bryk & Raudenbush, 1987, 1989, 1992; Cohen, 1988; Goldstein, 1987; Hox, 2002; Lee & Bryk, 1989; Raudenbush & Bryk, 1986, 2002; Raudenbush, Bryk, Cheong, & Congdon, 2001; Reise & Duan, 2002), structural equation modeling (e.g., Bollen, 1989; Bollen & Long, 1993; Byrne, 1998, 2001; Jöreskog & Sörbom, 1996a, 1996b; Loehlin, 2003), meta-analysis, and all sorts of other fancy models. Of particular interest for influencing policy decisions through contemporary data analysis is measuring temporal changes in targeted outcomes. This requires the use of growth-curve modeling (Bollen & Curran, 2005) to measure change in outcomes as a function of time and to predict the rate and pattern of growth. Growth-curve modeling of individual change circumvents the limitations inherent in traditional repeated-measures analysis of variance (Agresti & Finlay, 1997; Howell, 2007; Shadish, Cook, & Campbell, 2002), in which restrictive assumptions (such as sphericity or constant correlations over time) often are not met. Traditional growth-curve modeling ignores individual growth trajectories, which are treated as error, but has difficulty dealing with missing data and inconsistent time periods; these are severe problems because frequently longitudinal studies suffer from relatively high rates of attrition, and it is difficult to sustain repeated measurements at nearly equal intervals.

An enhancement of traditional growth-curve models is provided by the analysis of individual growth curves, in which within-individual change is modeled as a function of time, providing for both linear growth (instantaneous growth rate at intercept) and curvilinear growth (acceleration) in the level-1 (individual-level) model and with predictors of baseline performance, of initial learning rate, and of acceleration in the level-2 (aggregate) model. How to model the time variable is a major methodological issue (see van den Bergh et al., Chap. 20). For example, centering the time variable can dramatically change the interpretation of lower-order coefficients. Variance in the coefficients for individuals may reflect important individual differences in, for example, students' rate of learning and their sensitivity to contextual circumstances. Residuals from such individual growth-curve models reflect individual differences among students, which then can be used as predictors for other analyses. More complex models are possible by incorporating time-varying covariates at level 1, individual covariates at level 2, modeling heterogeneous level-1 variance and autocorrelation, and specifying complex error structures using a hierarchical, multivariate, linear model.

22.6 A Brief Segue

These and other dimensions of research methodology expectations rise to the forefront when the inevitable need arises to put together the research team that must consolidate qualitative and quantitative expertise with content knowledge and to develop the synergies that are essential for successful research proposals. To conclude this chapter, it may be helpful to take note of the practicalities of what must be done and the complexities that need to be addressed in the pursuit of funded research from a major grant opportunity directed toward impacting public policy.

To help make the preceding discussion about methodological research needs more concrete, we examine below the methodological requirements for IES request for proposal CFDA (Catalogue of Federal Domestic Assistance) 84.305A for the Education Research Grants program (US IES, 2008). All of these methodological requirements must be addressed, as specifically as possible. This requires teamwork between content and methods/measurement experts. The hoped-for result is a more vigorous, externally funded, research program.

22.6.1 Measurable Outcomes

(a) [r]eadiness for schooling (pre-reading, pre-writing, early mathematics and science knowledge and skills, and social development);
(b) [a]cademic outcomes in reading, writing, mathematics, and science;
(c) [s]tudent behavior and social interactions within schools that affect the learning of academic content;
(d) [s]kills that support independent living for students with significant disabilities; and
(e) [e]ducational attainment (high school graduation, enrollment in and completion of post-secondary education). (US IES, p. 8)

22.6.2 Five Research Goals

(a) Goal One – identify existing programs, practices, and policies that may have an impact on student outcomes and the factors that may mediate or moderate the effects of these programs, practices, and policies;
(b) Goal Two – develop programs, practices, and policies that are theoretically and empirically based …;
(c) Goal Three – establish the efficacy of fully developed programs, practices, and policies …;
(d) Goal Four – provide evidence on the effectiveness of programs, practices, and policies implemented at scale; and
(e) Goal Five – develop or validate data and measurement systems and tools. (p. 9)

22.6.3 Methodological Requirements

- Clear, concise hypotheses or research questions. (p. 51)
- Sample to be selected and sampling procedures to be employed …, including justification for exclusion and inclusion criteria. (p. 59) [Describe strategies to increase the likelihood that participants will remain in the study over the course of the evaluation (i.e., reduce attrition).]
- Detailed research design. … Studies using randomized assignment to treatment and comparison conditions are strongly preferred. … [C]learly state the unit of randomization (e.g., students, classroom, teacher, or school) … [and] explain the procedures for assignment of groups (e.g., schools, classrooms) or participants to treatment and comparison conditions. (p. 59)

Only when a randomized trial is not possible may alternatives that substantially minimize selection bias or allow it to be modeled be employed. Applicants proposing to use a design other than a randomized design must make a compelling case that randomization is not possible. Acceptable alternatives include regression-discontinuity designs or other well-designed, quasi-experimental designs that minimize the effects of selection bias on estimates of effect size through propensity score balancing or regression.

- The power of the evaluation design to detect a reasonably expected and minimally important effect … indicate clearly (e.g., including the statistical formula) how the effect size was calculated. (p. 60)

For clusters or groups of students randomly assigned to treatment and comparison conditions, consider the number of clusters, the number of individuals within clusters, the potential adjustment from covariates, the desired effect, the intraclass correlation (Killip, Mahfoud, & Pearce, 2004, i.e., the variance between clusters relative to the total variance between and within clusters), and the desired power of the design (note that other factors may also affect the determination of sample size, such as using one-tailed versus two-tailed tests, repeated observations, attrition of participants, etc.).

- Measures of student outcomes [including] researcher developed measures and … relevant … standardized measures of student achievement. (US IES, p. 61)
- Fidelity of implementation of the intervention … how the implementation of the intervention would be documented and measured. (p. 61)
- Compare intervention and comparison groups on the implementation of critical features of the intervention [to connect observed differences to treatment effects.] … [A]void contamination between treatment and comparison groups. (pp. 61–62)
- Mediating and moderating variables … that may explain the effectiveness or ineffectiveness of the intervention. … [A]ccount for sources of variation in outcomes across settings (i.e., to account for what might otherwise be part of the error variance). … [D]emonstrate the conditions and critical variables that affect the success of a given intervention. The most scalable interventions are those that can produce the desired effects across a range of education contexts. (p. 62)
- Data analysis. All proposals must include detailed descriptions of data analysis procedures. … Most evaluations of education interventions involve clustering of students in classes and schools and require the effects of such clustering to be accounted for in the analyses, even when individuals are randomly assigned to condition. Such circumstances generally require specialized multilevel statistical analyses using computer programs designed for such purposes. (p. 62)

22.7 Where Do We Go from Here?

The subsequent chapters contributed to Part V of this book span a wide variety of the implications for public policy of expectations/requirements for Gold Standards education research in many different countries and in diverse contexts. The next chapter examines the interplay between the needs for scientifically based research and the provision of research expenditures from the perspectives of education research in the United States, Canada, the European Union, Germany, the United Kingdom, Australia, New Zealand, southern Africa, and Taiwan (Republic of China). A similarly transnational range of views is afforded in the chapter on research ethics, which explores the diversity of policies in different countries and in different institutions regarding human subjects protections in education research and the varying extent to which constraints are imposed on education researchers. Another chapter addresses policies related to data sharing, including data disclosure, confidentiality, and security. Qualitative metasynthesis, applying aspects of meta-analysis from quantitative methodology, is addressed by another set of authors as a means for revealing general patterns in systematic research reviews, metasyntheses, secondary reanalyses, and case-to-case comparisons of qualitative research studies. The part concludes with a call to educators to make better use of the results of science to change social practice.

The chapters in this part, and indeed all the contributions throughout this book, reveal a compelling need for a self-conscious, deliberate, and directed effort to integrate methodology, policy, and advocacy into a coherent approach to speaking truth to power. Knowing what the truth is, when it is ripe for sharing, whom to share it with, and how to convey it with maximum impact are all essential aspects of what is to be done.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd edn.). Upper Saddle River, NJ: Prentice Hall.

American Statistical Association. (2007). *Using statistics effectively in mathematics education research*. Arlington, VA: Author. Available from http://www.amstat.org/research_grants/pdfs/SMERReport.pdf

Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement & Evaluation in Counseling & Development*, *25*(2), 58–90.

Best Evidence Encyclopedia. (n.d.). *Homepage*. Retrieved July 7, 2008, from http://www.bestevidence.org/

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective*. New York: Wiley.

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Brewer, D. J., & Goldhaber, D. D. (2008). Examining the incentives in educational research. *Phi Delta Kappan*, *89*(5), 361–364.

Briggs, D. C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher*, *37*(1), 15–22.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158.

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three level hierarchical linear model. In R. D. Bock (Ed.), *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming*. Mahwah, NJ: Lawrence Erlbaum.

Campbell Collaboration. (n.d.). *Homepage*. Retrieved October 3, 2007, from http://www.campbellcollaboration.org

Chatterji, M. (2008). Comments on Slavin: Synthesizing evidence from impact evaluations in education to inform action. *Educational Researcher*, *37*(1), 23–26.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum.

Cohn, D. (2006). Jumping into the political fray: Academics and policy-making. *Institute for Research on Public Policy (IRPP) Matters*, *7*(3), 8–36. Retrieved from http://www.irpp.org/pm/index.htm

Comprehensive School Reform Quality Center. (2006). *Homepage*. Retrieved July 21, 2008, from http://www.csrq.org/

Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London: Arnold.
Dynarski, M. (2008). Comments on Slavin: Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher*, *37*(1), 27–29.
Education Sciences Reform Act of 2002. Pub. L. No. 107–279, 116 Stat. 1940. (2002).
Evidence for Policy and Practice Information and Co-ordinating Centre. (n.d.). *Homepage*. Retrieved June 20, 2008, from http://eppi.ioe.ac.uk/cms/
Fusarelli, L. D. (2008). Flying (partially) blind: School leaders' use of research in decision making. *Phi Delta Kappan*, *89*(5), 365–358.
Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). *Reading first impact study: Interim report*. (NCEE 2008–4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Available from http://ies.ed.gov/ncee/pubs/20084016/index.asp
Goldstein, H. (1987). *Multilevel models in educational and social research*. New York: Oxford University Press.
Goldstein, H. (2003). *Multilevel statistical models* (3rd edn.). London: Hodder Arnold.
Green, J. L., & Skukauskaité, A. (2008). Comments on Slavin: Becoming critical readers: Issues in transparency, representation, and warranting of claims. *Educational Researcher*, *37*(1), 30–40.
Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational & Behavioral Statistics*, *32*(2), 151–179.
Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational & Behavioral Statistics*, *32*(4), 341–370.
Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation & Policy Analysis*, *29*(1), 60–87.
Henig, J. R. (2008). The evolving relationship between researchers and public policy. *Phi Delta Kappan*, *89*(5), 357–360.
Hess, F. M. (2008). The politics of knowledge. *Phi Delta Kappan*, *89*(5), 354–356.
Howell, D. C. (2007). *Statistical methods for psychology* (6th edn.). Pacific Grove, CA: Duxbury.
Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edn.). Thousand Oaks, CA: Sage.
Ingersoll, R. M. (2008). A researcher encounters the policy realm: A personal tale. *Phi Delta Kappan*, *85*(5), 369–371.
Jöreskog, K. G., & Sörbom, D. (1996a). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.
Jöreskog, K. G., & Sörbom, D. (1996b). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.
Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine*, *2*(3), 204–208.
Kim, J. S. (2008). Research and the reading wars. *Phi Delta Kappan*, *85*(5), 372–375.
Kingdon, J. W. (1995). *Agendas, alternatives, and public policies* (2nd edn.). New York: HarperCollins.
Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, *62*(3), 172–192.
Lenth, R. V. (2006). Java applets for power and sample size. Available from http://www.cs.uiowa.edu/~rlenth/Power/
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Loehlin, J. C. (2003). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th edn.). Mahwah, NJ: Lawrence Erlbaum.
Murphy, K. R., & Myors, B. (2003). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the national mathematics advisory panel*. Washington, DC: US Department of Education. Available from http://www.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf
No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425. (2002).
Ragin, C. C., Nagel, J., & White, P. (2004). Workshop on scientific foundations of qualitative research. Available from http://www.nsf.gov/pubs/2004/nsf04219/start.htm
Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185.
Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25–31.
Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd edn.). Thousand Oaks, CA: Sage.
Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2001). *HLM5: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International Inc.
Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213.
Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401.
Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation & Policy Analysis*, *29*(1), 5–29.
Raudenbush, S. W., Spybrook, J., Liu, X., Congdon, R., & Martinez, A. (2006). Optimal Design (Version 1.76) [Software]. Available from http://sitemaker.umich.edu/group-based/optimal_design_software
Reise, S. P., & Duan, N. (Eds.). (2002). *Multilevel modeling: Methodological advances, issues, and applications*. Mahwah, NJ: Lawrence Erlbaum.
Rotherham, A. J. (2008). The translators: The media and school choice research. *Phi Delta Kappan*, *89*(5), 376–379.
Schmitt, L. N. T., & Whitsett, M. D. (2008). Using evaluation data to strike a balance between stakeholders and accountability systems. In T. Berry & R. M. Eddy (Eds.), *Consequences of No Child Left Behind for educational evaluation* (pp. 47–58). San Francisco: Jossey-Bass.
Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. (Contract No. ED-01-CO-0038-0009; MPR Reference No. 6046-310). Princeton, NJ: Mathematica Policy Research, Inc. Available from http://www.mathematica-mpr.com/publications/PDFs/statisticalpower.pdf
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
Slavin, R. E. (2008a). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14.
Slavin, R. E. (2008b). Response to Comments: Evidence-based reform in education: Which evidence counts? *Educational Researcher*, *37*(1), 47–50.
Sloane, F. (2008). Comments on Slavin: Through the looking glass: Experiments, quasi-experiments, and the medical model. *Educational Researcher*, *37*(1), 41–46.
Toynbee, A. J. (1934–1961). *A study of history*. London: Oxford University Press.
Trenholm, C., Devaney, B., Fortson, K., Clark, M., Quay, L., & Wheeler, J. (2008). Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. *Journal of Policy Analysis and Management, 27*(2), 255–276.

United States Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available from http://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp

United States Institute of Education Sciences. (2008). *Education research grants CFDA number: 84.305A*. Retrieved June 17, 2008, from http://ies.ed.gov/funding/pdf/2009_84305A.pdf

United States Institute of Education Sciences. (n.d.-a). *Homepage*. Retrieved June 17, 2008, from http://ies.ed.gov/

United States Institute of Education Sciences. (n.d.-b). *What Works Clearinghouse: Homepage*. Retrieved June 20, 2008, from http://ies.ed.gov/ncee/wwc/

United States Institute of Education Sciences. (n.d.-c). *What Works Clearinghouse: Technical assistance*. Retrieved July 8, 2008, from http://ies.ed.gov/ncee/wwc/tech_assistance/

United States Institute of Education Sciences. (n.d.-d). *What Works Clearinghouse: Technical working papers*. Retrieved July 8, 2008, from http://ies.ed.gov/ncee/wwc/twp.asp

United States National Research Council. (2002). *Scientific research in education*. Committee on Scientific Principles for Education Research. R. J. Shavelson & L. Towne (Eds.). Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

W. T. Grant Foundation and University of Michigan (n.d.). *Building capacity to evaluate group-level interventions*. Retrieved July 21, 2008, from http://sitemaker.umich.edu/group-based/home

Wildavsky, A. (1979). *Speaking truth to power: The art and craft of policy analysis*. Boston: Little, Brown.

Ye, N. (Ed.). (2003). *The handbook of data mining*. Mahwah, NJ: Lawrence Erlbaum.