

1 Public Good and Private Interest in Educational Evaluation

Sandra Mathison

If we share an apple, we can share a community

So reads the caption on a poster I received upon completion of an evaluation of an accelerated secondary school program. Over a six-month period I worked with this school and its stakeholders (including students, teachers, parents, school administrators, and graduates) to evaluate this accelerated program for precocious eighth and ninth graders. The evaluation was stakeholder based, the types of evidence collected evolved as the evaluation proceeded, the students became interested in the process and worked with me to develop data collection skills and then collect data themselves. The final report and where to go from the results was a collaborative effort among the stakeholder groups.

This is typically the kind of educational evaluation I do, although there is a wide range of approaches used in educational evaluation. But for this work, I did not get paid. The school has no money for evaluation. The school district has no money for evaluation. The evaluation approach the school wanted is not a priority or publicly funded.

Consider this quote from the *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*, published by the U.S. Department of Education (2003):

Well-designed and implemented randomized controlled trials are considered the “gold standard” for evaluating an intervention’s effectiveness, in fields such as medicine, welfare and employment policy, and psychology. This section discusses what a randomized controlled trial is, and outlines evidence indicating that such trials should play a similar role in education.

And so the stage for the current focus on educational evaluation is set by the U.S. Department of Education.

Educational Evaluation in Historical Context

Educational evaluation is by and large a public good—although evaluation occurs in many fields and in many contexts supported through many means, the genesis of educational evaluation is the stipulations in the Elementary and Secondary Education Act (ESEA) passed in 1965. Established as part of Lyndon Johnson’s War on Poverty, the ESEA provides federal assistance to schools, communities, and children in need. With current funding of about \$9.5 billion annually, the ESEA continues to be the single largest source of federal funding to K-12 schools. Through its many Title programs, especially Title I, ESEA has been a major force in focusing how and what is taught in schools, as well as

6 *Sandra Mathison*

the ways those activities are evaluated. With Johnson's conceptualization of ESEA, educational evaluation was seen to be a public good (just like education and schooling) that should serve the common public good. What I want to illustrate is that although educational evaluation remains a public good it increasingly serves private interests.

While the passage of ESEA marks the beginning of the formalization of educational evaluation, one prior event, the Eight Year Study, also played an important role in educational evaluation, although it is more often associated with developments in curriculum theory and design. The Eight Year Study involved 30 high schools dispersed throughout the United States and serving diverse communities (Aiken, 1942). Each school developed its own curriculum and each was released from government regulations, as well as the need for students to take college entrance examinations. With dissension early in the project about how its success should be evaluated, a young Ralph Tyler was brought on board to direct the evaluation, which was funded by the Rockefeller Foundation. Out of the Eight Year Study came what is now known as the "Tyler Rationale," the common-sense idea that what students were supposed to learn should determine what happened in classrooms and how evaluation should be done (Tyler, 1949).

Tyler's evaluation team devised many curriculum specific tests, helped to build the capacity for each school to devise its own measures of context specific activities and objectives, identified a role for learners in evaluation, and developed data records to serve intended purposes (including descriptive student report cards) (Smith & Tyler, 1942). All of these developments resonate with conceptual developments in evaluation from the 1970s to the present. The notion of opportunity to learn is related to the curriculum sensitivity of measures; the widespread focus on evaluation capacity building resonates with the Tylerian commitment to helping schools help themselves in judging the quality and value of their work; democratic and empowerment approaches, indeed all stakeholder based approaches, resonate with the learners' active participation in evaluation; and the naturalistic approaches to evaluation resonate with the use of behavioral descriptive data.

The Eight Year Study ended in 1941 and was published in five volumes in 1942, an event which was overshadowed by its unfortunate coincidence with U.S. troops taking an active role in World War II. Nonetheless, Ralph Tyler and the Eight Year Study evaluation staff provided a foundation, whether always recognized or not, for future education evaluators.

When ESEA was passed in 1965 (legislation which Ralph Tyler had a hand in as an educational advisor to the Johnson administration) the requirement that the expenditure of public funds be accounted for thrust educators into a new and unfamiliar role. Educational researchers and educational psychologists stepped in to fill the need for evaluation created by ESEA. But the efforts of practitioners and researchers alike were generally considered to be only minimally successful at providing the kind of evaluative information envisioned. The compensatory programs supported by ESEA were complex and embedded in the complex organization of schooling.

Since the federal politicians, especially ESEA architect Robert Kennedy, were primarily interested in accountability, evaluation requirements for ESEA, especially for Title I, which was the largest compensatory program, emphasized uniform procedures and comparable data at the state and national levels, a direction many evaluators found misdirected (Jaeger, 1978; Wiley, 1979). During this period, the advances in educational evaluation were, at least in part, over and against the federal approach to evaluation, especially Title I programs, primarily a focus on student achievement (expressed as "normal curve equivalents").

There is of course nothing wrong with knowing how well or how poorly a student performs. Yet schools, insofar as they are educational institutions, should not be content with performance. Education as a process is concerned with the cultivation of intellectual power, and the ability to determine what a student knows is not necessarily useful or sufficient for making the process more effective. (Eisner, 1979, p. 11)

The tension between meeting federally mandated reporting requirements and local needs for evaluative information was a significant part of the debate. School districts did the minimum to meet the federal reporting guidelines, but at the same time often looked for guidance as to how to sincerely evaluate what was happening in local schools. While school districts may have needed only one person to meet the reporting mandate, a broader local interest led to the creation of evaluation departments in many, and certainly all of the large, school districts.

The late 1960s and into the 1980s were the gold rush days of educational evaluation. During this time, models of evaluation proliferated and truly exciting intellectual work was being done, especially in education. Often very traditionally educated educational psychologists experienced epiphanies that redirected their thinking about evaluation. For example, Robert Stake, a psychometrician who began his career at the Educational Testing Service, wrote a paper, "Countenance of Educational Evaluation," which reoriented thinking about the nature of educational interventions and what was important to pay attention to in determining their effectiveness (Stake, 1967). Egon Guba, a well known education change researcher abandoned the research, development, diffusion approach for naturalistic and qualitative approaches that examined educational interventions carefully and contextually (Lincoln & Guba, 1985). Lee Cronbach, psychometric genius, focused not on the technical aspects of measurement in evaluation, but on the policy-oriented nature of evaluation. This was an idea that led to a radical reconstruction of internal and external validity, which included separating the two conceptually and conceptualizing external validity in relation to usability and plausibility of conclusions, not as a technical feature of research or evaluation design (Cronbach, 1982).

While ESEA, now No Child Left Behind (NCLB), is the driving force in what counts as evaluation in education, other developments have occurred simultaneously and are important companion pieces to understanding the contemporary educational evaluation landscape. The National Assessment of Educational Progress (NAEP), sometimes referred to as the nation's report card, was created in the mid-1960s, building on and systematizing a much longer history of efforts to use educational statistics to improve and expand public education. Francis Keppel, the U.S. Commissioner of Education from 1962 to 1965 and a former dean of the Harvard School of Education, lamented the lack of information about the academic achievement of American students.

It became clear that American education had not yet faced up to the question of how to determine the quality of academic performance in the schools. There was a lack of information. Without a reporting system that alerted state or federal authorities to the need for support to shore up educational weakness, programs had to be devised on the basis of social and economic data.... (Keppel, 1966, pp. 108–109)

Under the direction of Ralph Tyler (Tyler's intellectual legacy in evaluation is huge, as noted by his continued involvement in pivotal events in educational evaluation), NAEP developed as a system to test a sample of students on a range of test items, rather than the simple testing of all students with the same test items. To allay fears that NAEP would

be used to coerce local and state educational authorities the results were initially released for four regions only. NAEP has continued to develop, early on largely with the use of private funding from the Carnegie Corporation, and the early fears of superintendents and professional associations (such as the National Council for Teachers of English) turn out to be well founded. State level NAEP scores are indeed now available. This shift in the use of NAEP occurred during the Reagan administration with then Secretary of Education Terrel Bell's infamous wall chart. With a desire to compare states' educational performance, indicators available for all states were needed, and NAEP filled that bill. Southern states, such as Arkansas, under then governor Bill Clinton, applauded the use of such comparisons believing that they would encourage competition, a presumed condition for improvement.

During these halcyon years in educational evaluation, much evaluation was publicly funded, primarily by the U.S. Department of Education, but also by other federal agencies such as the National Science Foundation, in addition to many foundations, such as Carnegie, Rockefeller, Ford, and Weyerhaeuser. The dominance of public money and the liberal and progressive political era contributed significantly to the conceptualization of evaluation as a public good. Discussions of how best to judge if education and schooling are good contributed to a lively national debate about what counts as good education and schooling.

For example, the relatively small number of meta-evaluations conducted during this time focused primarily on whether the evaluation was fair and in the public interest. Two good examples of this are the meta-evaluation of Follow-Through (House, Glass, McLean, & Walker, 1978) (that thoroughly criticized economist Alice Rivlin's planned variation experiment as an evaluation method that did not do justice to the unique contributions of follow-through models for local communities) and the meta-evaluation of Push-Excel (House, 1988), Jesse Jackson's inspirational youth program that was undone by Charles Murray's evaluation that failed to consider the program on its own terms in the context of local communities (Murray was coauthor with Richard Herrnstein of *Bell Curve: Intelligence and Class Structure in American Life*).

The New Neoliberal Era and Educational Evaluation

The reauthorization of ESEA, now called No Child Left Behind (NCLB), reinforces the need for evaluation. But unlike the more general expectation for evaluation that typified the original ESEA evaluation mandate, NCLB is decidedly more prescriptive about how education should be evaluated, in part because of the inclusion of sanctions for failure to perform. While earlier versions of ESEA focused on student performance, NCLB did so by invoking the notion of "annual yearly progress" (AYP) and continued funding from the federal government was now dependent on each school making "continuous and substantial progress" toward academic proficiency by *all* students.

The 1965 authorization of ESEA, in spite of its emphasis on uniformity and standardization, opened new frontiers and contributed significantly to the discipline of evaluation, but NCLB has narrowed the scope of evaluation. Fewer federal funds are now spent on educational evaluation and the burden of evaluation has been shifted to the state and local levels through student testing. NCLB mandates what counts as evaluation (acceptable indicators, what counts as progress, consequences for lack of progress) but provides no funding to carry out the mandate. George W. Bush declared that with the reauthorization of NCLB, "America's schools will be on a new path of reform, and a new path of results." No one would disagree. Teaching has become less professional and more mechanical (Mathison & Freeman, 2003); business and profits for the test publishing and

scoring companies has increased markedly, even though the testing is mostly misdirected (Popham, 2004); and schools chase unattainable goals out of fear (Linn, 2005). That the path is new does not necessarily mean it is the path best traveled.

The current narrow evaluation focus of NCLB (standardized tests for evaluating student learning and schools) evolved as a result of changes in political values. The current public and governmental neoliberalist sentiment (an ideology shared by Republicans and Democrats) has had major implications for government policies beginning in the 1970s but increasingly prominent since 1980. Neoliberalism deemphasizes government intervention in the economy, focusing instead on achieving progress (including social justice) by encouraging free market methods and fewer restrictions on business operations and economic development.

Concerns about and constructions of a crisis in American schools are formulated around constructs such as international competitiveness and work productivity. In other words, our schools are meant to serve the interests of the economy. *A Nation At Risk*, published in 1983, was the clarion call for educational reform: “The educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people.... We have, in effect been committing an act of unthinking, unilateral educational disarmament.”

Although it took a few years, in 1989 President Bush and the state governors called an Education Summit in Charlottesville. That summit established six broad educational goals to be reached by the year 2000. President Clinton signed Goals 2000 into law in 1994. Goals 3 and 4 were related specifically to academic achievement and thus set the stage for both what educational evaluation should focus on and how.

In 1990, the federally funded procedures for moving the country toward accomplishment of these goals were established. The National Education Goals Panel (NEGP) and the National Council on Education Standards and Testing (NCEST) were created and charged with answering a number of questions: What is the subject matter to be addressed? What types of assessments should be used? What standards of performance should be set?

In 1996, a national education summit was attended by 40 state governors and more than 45 business leaders. They supported efforts to set clear academic standards in the core subject areas at the state and local levels and the business leaders pledged to consider the existence of state standards when locating facilities. Another summit followed in 1999 and focused on three key challenges facing U.S. schools—improving educator quality, helping all students reach high standards, and strengthening accountability—and agreed to specify how each of their states would address these challenges. And a final summit occurred in 2001, when state governors and business leaders met at the IBM Conference Center in Palisades, New York to provide guidance to states in creating and using tests, including the development of a national testing plan. The culminating event to this series of events beginning in the early 1980s was the passage of NCLB.

The heavy hand of business interests and market metaphors in establishing what schools should do and how we should evaluate what they are doing is evident in the role business leaders have played in the education summits. The infrastructure that supports this perspective is broad and deep. The Business Roundtable, an association of chief executive officers of U.S. corporations, and the even more focused Business Coalition for Education Reform, a coalition of 13 business associations, are political supporters and active players in narrowing evaluation of education to the use of standardized achievement tests.

With the passage of NCLB, the U.S. Department of Education simultaneously funds less evaluation. This is partly because of a much-narrowed definition of what the government

now considers good evaluation and partly because the U.S. Department of Education sees itself as the judge of educational evaluation and research, rather than its sponsor.

The U.S. Department of Education recognizes four kinds of program evaluation: (1) continuous improvement (employing market research techniques); (2) program performance data (use of performance based data management systems); (3) descriptive studies of program implementation (use of passive, descriptive techniques like surveys, self-reports, and case studies); and (4) rigorous field trials of specific interventions (field trials with randomized assignment). It is this last sort of evaluation that is the *pièce de résistance*, what are referred to as the “new generation of rigorous evaluations.” It is this evaluation approach that permits entry to the What Works Clearinghouse (WWC) of the Institute of Education Sciences (IES), and thus an intervention, practice, or curriculum earns the governmental imprimatur of an “evidence based best practice.”

As reflected in the quote at the beginning of this chapter, evaluations must be randomized clinical trials, perhaps quasi-experimental or regression discontinuity designs. Few if any educational evaluations have been of this sort, indeed much of the work since the 1960s has been directed to creating different evaluation methods and models of evaluative inquiry (not just borrowed research methods) that answer evaluative questions. Questions about feasibility, practicability, needs, costs, intended and unintended outcomes, and ethics, and justifiability are uniquely evaluative.

While neoliberalism clearly surrounds NCLB by the characterization of education as a commodity, the use of single indicators, and the promotion of market systems to improve the quality of schooling, the connection to the U.S. government mandate for randomized clinical trials is a little more tenuous. However, neoliberalism is characterized by a reliance on specialized knowledge and silencing or at least muting the voices of the populace. Unlike many approaches to evaluation that are built on the inclusion of stakeholders in directing and conducting the evaluation, experimental design is controlled by experts, and stakeholders (especially service providers and recipients) are conceived of more as anonymous subjects, and less as moral, sociopolitical actors.

By many accounts, the discipline of evaluation dealt with the role of experimental design in evaluation—it is potentially useful, most of the time impractical, and often limited in answering the array of evaluative questions invariably asked. It was unclear that the deep commitment to experimental design as the *sine qua non* of evaluation designs was only dormant, waiting for the political fertilizer to germinate and grow this commitment.

Evaluation Foci and Methods that Encourage Private Interests

Just as progressivism was the value context up to the late 1970s and even early 1980s, neoliberalism has been the value context that brings educational evaluation to where we are today. Schools are a business, education is a product, products should be created efficiently, and one should look to the bottom line in making decisions. Implicit in this neoliberal perspective are values (and rhetoric) that motivate action. The most obvious of these values is that accountability is good, that simple parsimonious means for holding schools accountable are also good, that choice or competition will increase quality, that it is morally superior to seek employability over other purposes of education. Econometrics drives thinking about what these simple parsimonious means are—the appeal of single indicators like standardized tests and the concept of value added now promoted in evaluating the teachers.

It is useful to look at two examples that illustrate this neoliberal focusing of evaluation.

Example 1: SchoolMatters

SchoolMatters describes its purpose thus: “SchoolMatters gives policymakers, educators, and parents the tools they need to make better-informed decisions that improve student performance. SchoolMatters will educate, empower, and engage education stakeholders.”

It is a product of Standard and Poors, which is in turn owned by McGraw-Hill Companies—the biggest producer of educational tests; and it promises to provide, in one convenient location (schoolmatters.com) the following:

- Student Performance: national and state test results, participation, attendance, graduation, and dropout–promotion rates.
- Spending, Revenue, Taxes, and Debt: financial data for each school district, along with state and county comparisons.
- School Environment: class size, teacher qualifications, and student demographics.
- Community Demographics: adult education levels, household incomes, and labor force statistics.

And the highly interactive website delivers on this. A range of indicators are used: school size, reading scores, math scores, special needs (limited to information about English Language Learners), teacher–student ratios, ethnicity, income, and housing costs. But there is much about schools and education that SchoolMatters does not deliver, because it is considered unnecessary, or difficult to collect or aggregate data, or does not reflect a narrow conception of the purpose of schools to prepare skilled workers. Indicators you will not find: types of school programming, health and fitness, quality of physical plant, availability of resources such as books/paper/pencils, attrition rates, proportion of dropouts getting a GED, and levels of volunteerism or community involvement.

In addition, there is a decidedly different language used in the discussion of factors that are outside of school control versus those that are in school control. In the former case there are cautions about the importance of parents and communities in academic achievement.

Research has shown that the education levels and contributions of parents are critical factors that impact a child’s academic performance. To help all students reach their full potential, it is necessary that students, teachers, families, and communities collectively engage in efforts to improve student performance.

The implication here is that parents should get themselves educated and do something to contribute to the improvement of student performance—an essentially moral message to others.

This contrasts with a factor that is within the school’s control, namely class size. When there is a potential change to the school that might improve student performance but at a cost, SchoolMatters advises caution.

Smaller class sizes may improve student performance in certain settings; for example, research has shown that low-income students in early grades may benefit from smaller classes. Yet, there is less agreement on across-the-board benefits of small classes. Deciding to implement a policy to create smaller classrooms means that more teachers must be hired, and not all communities have a pool of qualified teachers from which to draw.

This selective presentation of research on the benefits of reducing class size serves other purposes—it diverts parents or educators from investing time in contemplating changes that might increase costs and therefore threaten the production of educational products at the lowest possible cost. Indeed, S&P promotes “improving the return on educational resources” rather than insuring there are adequate resources.

Example 2: What Works Clearinghouse

As mentioned earlier, the What Works Clearinghouse (WWC) promises “a new generation of rigorous evaluations,” by specifying a single acceptable, desirable evaluation design, the randomized clinical trial (RCT). The WWC means to “identify studies that provide the strongest evidence of effects: primarily well conducted randomized controlled trials and regression discontinuity studies, and secondarily quasi-experimental studies of especially strong design.”

With fewer resources for funding educational evaluation, the U.S. Department of Education has turned its attention via the WWC to evaluating research and evaluation.

The WWC aims to promote informed education decision making through a set of easily accessible databases and user-friendly reports that provide education consumers with high-quality reviews of the effectiveness of replicable educational interventions (programs, products, practices, and policies) that intend to improve student outcomes.

The WWC standards for identifying studies that show what works are based on an examination of design elements only. The conclusions of studies that meet the design standards are those worth paying attention to, regardless of their substance.

Evaluation in Service of What?

These two examples demonstrate different but complementary ways that evaluation as a public good has come to serve private interests. Metaphorically, a database for rating schools created by a market rating firm implicitly reinforces and naturalizes the neoliberal market oriented values that inform what schooling is and how we evaluate it. And the narrowly defined and highly specialized conception of evaluation design promoted by the Institute for Education Science and manifest in the What Works Clearinghouse delineate what counts as evaluative knowledge and thus what counts as worth knowing about education and schooling.

These are powerful forces, and evaluators working within public school districts across the nation would agree they do little “evaluation” any more. They are too busy with NCLB’s mandates for standardized testing of students and trying to figure out how, if at all, they can conjure up an RCT so as to obtain much needed money for local evaluation efforts (Mathison & Muñoz, 2004).

Educational evaluation has been and will continue to evolve. In its early days educational evaluation bore the mark of progressivism. Education and the evaluation of it were supported by public funding and defined as a public good, in the interest of all. Evaluation reflected these values (including efficiency, social justice, and democracy) and was financially supported with public funds. In the 1980s the values of progressivism gave way to the emerging values of neoliberalism. All evaluation requires the specification of the good-making qualities of what is being evaluated. And these good-making qualities are socially constructed; therefore the dominant approach to education evalu-

ation reflects the current sociopolitical zeitgeist. In the current state of neoliberalism and neoconservatism, education and the evaluation of it increasingly reflect those values, including commodification, privatization, and Judeo-Christian morality. The practice of evaluation (like many social practices) is a reflection of the values, beliefs, and preferences of the time.

Some educational evaluators still follow in the footsteps of Tyler, Stake, Guba, and Cronbach and, although in small measure, continue to conduct evaluations that encourage broad understandings of quality in education and schooling. Participatory and stakeholder based approaches to educational evaluation can be found, approaches that recognize the importance of local context, the critical involvement of especially those who are most powerless in judgments of education supposedly in their best interests, the need to consider the complex purposes of schooling that include basic skills for employability but extend beyond to social development and citizenship (Mathison, in press). Evaluation capacity building, a strategy for helping people make their own judgments about the quality of their schools, is an underlying goal for a number of evaluators (King, 2005). And grassroots groups, such as the Massachusetts Coalition for Authentic Reform in Education (MassCARE), resist the narrow and singular criterion of academic achievement in considering the quality of education. There is still educational evaluation that builds on the explosive and exciting period of the 1970s and 80s—I still do it, but sometimes I get paid in posters.

References

- Aiken, W. M. (1942). *The story of the Eight Year Study*. New York: Harper Brothers. Retrieved February 13, 2007, from <http://www.8yearstudy.org>
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.
- Eisner, E. W. (1979). The use of qualitative forms of evaluation for improving educational practice. *Educational Evaluation and Policy Analysis*, 1(6), 11–19.
- Herrnstein, R. & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
- House, E. R. (1988). *The politics of charisma: The rise and fall of PUSH/Excel Program*. Boulder, CO: Westview Press.
- House, E. R., Glass, G. V., McLean, L. D., & Walker, D. C. (1978). No simple answer: A critique of the follow-through evaluation. *Harvard Educational Review*, 48, 128–160.
- Jaeger, R. M. (1978). *The effect of test selection of Title I project impact*. Paper presented at the annual meeting of the American Educational Research Association, Toronto.
- Keppel, F. (1966). *The necessary revolution in American education*. New York: Harper & Row.
- King, J. A. (2005). A proposal to build evaluation capacity at the Bunche-Da Vinci Learning Partnership Academy. In M. C. Alkin & C. A. Christie (Eds.), *New directions for evaluation*, 106, 85–98.
- Lincoln, Y., & Guba, E. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, 13(33). Retrieved February 13, 2007, from <http://epaa.asu.edu/epaa/v13n33/>
- Mathison, S. (in press). Serving the public interest through educational evaluation: Salvaging democracy by rejecting neo-liberalism. In C. B. Cousins & K. Ryan (Eds.), *Handbook of educational evaluation*. Newbury Park, CA: Sage.
- Mathison, S., & Freeman, M. (2003). Constraining the work of elementary teachers: Dilemmas and paradoxes created by state mandated testing. *Education Policy Analysis Archives*, 11(34). Retrieved February 13, 2007, from <http://epaa.asu.edu/epaa/v11n34>

14 Sandra Mathison

- Mathison, S., & Muñoz, M. (2004). Evaluating education and schools for democracy. In S. Mathison & E. W. Ross (Eds.), *Defending public schools: The nature and limits of standards based educational reform and testing* (pp. 71–81). Westport, CT: Greenwood Press.
- Nation at Risk (1983). [Http://www.ed.gov/pubs/NatAtRisk/index.html](http://www.ed.gov/pubs/NatAtRisk/index.html)
- Popham, J. (2004). Standards based education: Two wrongs don't make a right. In S. Mathison & E. W. Ross (Eds.), *Defending public schools: The nature and limits of standards based educational reform and testing* (pp. 1–26). Westport, CT: Greenwood Press.
- Smith, E. R., & Tyler, R. W. (1942). *Appraising and recording student progress*. New York: Harper.
- Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record*, 68(7), 523–540.
- Tyler, R. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.
- U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved February 12, 2007, from <http://www.ed.gov/rschstat/research/pubs/rigorousetid/index.html>
- Wiley, D. (1979). Evaluation by aggregation: Social and methodological biases. *Educational Evaluation and Policy Analysis*, 1(2), 41–45.