Presidential Address

Rebecca A. Maynard[1]

Evidence-Based Decision Making: What Will It Take for the Decision Makers to Care?

Over the past 10 years, there has been a growing emphasis on evidence-based policy and practice throughout in the United States and elsewhere around the world. This should be music to the ears of the APPAM membership, as most of us entered our professions, in no small part, to feed our personal interests in helping make the world a better place. The field of public policy research has changed enormously in the 26 years since the founding of APPAM in ways that have improved the quality of evidence that can be produced, expanded the breadth of questions addressed, and broadened avenues for communicating findings. Still, we are far from a world in which evidence is routinely and smartly produced and integrated into decision making.

APPAM's mission is to "[improve] public policy and management by fostering excellence in research, analysis, and education." Yet, much of the research we produce is ignored or misused for three main reasons. First, some of the research produces results that lack credibility. We know how to judge the credibility of study findings and should be vigilant in doing so. Second, much research addresses questions that have intrinsic interest and value, but yield results that are not helpful to policymakers and practitioners and, indeed, may be misinterpreted by them. We should sift and sort the information we direct to the policy and practitioner communities to promote proper interpretation and application of evidence. Third, there are many terrific examples of valuable syntheses of evidence on particular issues. However, there also are many examples of nonsystematic (and, in some cases, biased) reviews of evidence. There are well-established, but poorly disseminated methods for systematically synthesizing evidence on particular questions in ways that provide clear guidance regarding what we do and do not know with a particular degree of confidence.

Three personal experiences that have occurred during the course of trying to provide policy-relevant answers to important questions illustrate some challenges in using research for the development and management of public policy and to guide practice.

Experience 1. The question is: *Are there things that could or should be done to improve the neighborhoods where children live?* Three evidence-based answers are: *Yes, No, and Maybe.*

[1] This talk reflects much that I have learned from my colleagues in APPAM and elsewhere, including those who have been my collaborators, my teachers, my mentors, and my students. There are too many of you to name individually. But you know who you are. One or more of you deserves credit for any true insight that may be lurking in this talk. I am grateful to Phoebe Cottingham, Stuart Kerachsky, Lauren Scher, and Matthew Stagner for very useful comments on an early draft of this talk. I alone am responsible for any errors or omissions.

Journal of Policy Analysis and Management, Vol. 25, No. 2, 249–265 (2006)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20169

Recently, I was charged with what was an enriching, but challenging, task—to read and synthesize a body of "cutting-edge" research on a topic outside my area of expertise—the role of neighborhoods on developmental paths and outcomes for youth. The opening line in my 10-minute presentation on the research was: "Neighborhoods matter, but effects are modest and not as big as the influence of family, peers, or schools."

Before the ink had dried on my opening line, a newly released policy brief entitled *Overcoming Concentrated Poverty and Isolation* arrived on my desk. This brief proclaimed that "living in these high-poverty communities undermines the long-term life chances of families and children . . ." (Turner & Rawlings, 2005). After double checking my facts, I stuck by my opener. However, one member of my audience responded to my presentation with puzzlement, having recently heard two leading social scientists present a paper concluding that "neighborhoods don't matter."

Example 2. The question is: *Are pregnancy prevention programs effective in delaying sexual debut, in reducing pregnancy risk, and in reducing pregnancy?* Three evidence-based answers are: *Yes, No, and Sometimes, but we can't forecast which programs will be effective for whom, or under what conditions*.

Context: Three years ago, I embarked on a collaborative effort to systematically review the evidence related to this question (Scher, Maynard, & Stagner, 2005). Prior reviews of the research had drawn a range of conclusions. For example, one review of 75 studies based on experimental and nonexperimental methods concluded that "professionals working with youth . . . should replicate those programs that have the best evidence for success" (Kirby, 2000). According to the review, such programs had 10 common features. In contrast, another review that included 26 randomized controlled trials concluded, "We do not have a clear solution to the problem of high pregnancy rates among adolescents" (DiCenso, Guyatt, Willan, & Griffith, 2002).

We spent three years struggling with issues of what constitutes credible evidence and how best to synthesize findings across studies that vary along many dimensions—the intervention, the target population, and the setting, to name a few. In the end, we concluded that most studies did not answer the questions relevant to our review and/or did not yield credible answers to those questions. Among the studies that were reasonably well done and that addressed the right questions, relatively few showed evidence that the programs delayed sex, reduced pregnancy risk, or reduced pregnancy. Even when evidence of favorable impacts exists, the impacts tend to be small. Moreover, the studies show evidence that programs produced undesirable consequences much more often than can be accounted for by chance. It is important to note that there also are no consistent predictors of which types of programs will have favorable impacts and which will not.

Example 3. The question is: *Will giving parents vouchers that can be used to defray costs of sending their children to private schools lead to improved educational outcomes for children?* Three evidence-based answers are: *a little for African American children; a little for children who have only one African American parent; and probably not*.

Recently, a student who is conducting a systematic review of evidence on the effects on student outcomes of school vouchers came to class with her stack of articles, including several studies of the New York City School Voucher Program. Her question was: "How do I know which results to include in my review, as they differ?" The results of the original study of the New York City School Voucher Program showed evidence of impacts for African American youth only (Myers, Peterson, Mayer, Chou, & Howell, 2000; and Mayer, Peterson, Myers, Clark Tuttle, & Howell,

2002). In contrast, a re-analysis of the study data (retaining a large group of kindergarteners who were left out of the original analysis) shows no benefits of vouchers except for the small subset of students who have only one African American parent (Krueger & Zhu, 2004). In this and the several other cases presented by the student, we embarked on a tedious process of chronicling and judging the assumptions and analytic methods underlying the various estimates.[2]

When those of us who produce much of the evidence intended to guide economic, social, and education policy cannot agree on the answers to questions like these, is it any wonder that decision makers do not routinely turn to us as the source of truly reliable advice? If we are serious about wanting to increase the reliance by decision makers on our work for the betterment of society, we need to provide them with information that is relevant, that is reliable, and that is synthesized and presented smartly. Yet, there are at least three factors that complicate this—the relevant questions for policy and practice decisions change constantly; what constitutes reliable evidence on one question can be unreliable evidence on another; and accessibility of evidence is affected by the presentation of findings from individual studies, as well as whether and, if so, how the evidence is accumulated across studies.

KNOWING WHAT EVIDENCE IS NEEDED TO INFORM POLICY, MANAGEMENT, AND PRACTICE

It is rarely, if ever, possible to determine, a priori, exactly what evidence will be most useful to guide public policy or practice. Indeed, it would be more realistic to think about evidence-based policy and practice as a process. The set of relevant questions change over time; an accumulation of evidence generally is necessary to have a major impact on policy; and social, economic, and political trends alter the policy agenda in important ways. In addition, more often than not, the relevant questions and their answers are both complex and sensitive to context.

The important questions change over time. The quite dramatic way questions evolve is illustrated by trends related to employment and training policies over the past 40 years. Prior to the 1970s, for example, employment and training policy concerns were focused mainly on adult men, who were the primary "breadwinners" in American families.[3] As single parenthood became increasingly common, however, the focus shifted to encompass concerns about whether and, if so, how to promote better employment outcomes for single mothers who were likely to lose eligibility for public assistance as their children reached adulthood. By the mid-1980s, there was widespread sentiment that single mothers with no preschool-age children should be expected to work and contribute to the financial support of their families. And, by the late 1980s and early 1990s, there was increasing support for reducing public responsibility for long-term economic assistance to able-bodied adults and their families—a change that stimulated efforts to better prepare youth and young adults, in particular, for the labor market.

It is important to accumulate evidence. Rarely will any one study provide all the information needed to guide policy or practice in particular areas. Most often, the information needed to make well-informed decisions comes from multiple

[2] The same day we began struggling to understand the two sets of results from the New York City School Voucher Program, *the Wall Street Journal* reported on the dispute between Jesse Rothstein and Caroline Hoxby over another approach to examining the benefits of school choice for children—an approach that generates opposite conclusions depending on some untestable assumptions underlying the analysis (Hilsenrath, 2005).

[3] O'Connor (2002) provides an excellent history of the public assistance and related employment policies.

(often many) studies that examine issues in varied contexts and with different target populations. One of the most notable examples of the value added from persistent inquiry relates to the welfare reforms that were instituted in the mid-1990s after many years of experimental research to test a wide range of policies and practices. The earliest tests of welfare policy interventions focused on specific, somewhat idealistic, interventions. An example is the Supported Work Demonstration, which tested the effectiveness of supported, graduated-stress employment as a means for moving highly disadvantaged groups of unemployed youth and adults into full-time employment and self-sufficiency. This study examined programs serving four quite different target populations—long-term welfare recipients, ex-addicts, ex-offenders, and young school dropouts—in 10 urban settings (Hollister, Kemper, & Maynard, 1984).[4] The programs benefited only long-term welfare recipients, and even those impacts were modest.

The many subsequent rigorous field tests of strategies to promote and/or sustain self-sufficiency among low-income individuals entailed increased collaboration with policy and program partners on the design of the nature of the policies and practices to be tested and on the specification of the important questions to be addressed in the evaluations (Greenberg & Shroder, 2004). Studies were conducted in different state policy contexts, in different economic environments, and with different target populations. Indeed, there was somewhat of a "hunt and peck" strategy for identifying effective policies and practices to promote economic security among unemployed and low-wage workers. Again, the findings were predominantly null or disappointingly modest benefits.

It is important, however, that this accumulation of evidence was very instrumental in the welfare reforms that followed the passage of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA) and, to a lesser degree, in the design of the Work Force Investment Act of 1998.[5] Decision makers wanted to see replications of findings, consistent and interpretable patterns of findings, and findings that could be explained in ways that relate to what we know about human behavior. Consequently, the ultimate importance of the research derived as much from the correlational evidence relating patterns of program impacts (the causal evidence) with program features, population characteristics, and/or other contextual factors (Gueron & Pauly, 1991; and Mead, 1997).

This experience taught us that the types of interventions we study are unlikely to be the recommended policies or practices that ultimately are adopted. It also taught us the value of amassing evidence from studies examining similar or related questions in different contexts, with different intensities, and focused on different target populations. What ultimately influenced policy in a big way was the accumulation of evidence showing that: (1) small improvements in economic well-being could be achieved through changes in policies that required work-directed activity; (2) there were no measured adverse effects of such policies for parents or for their children; and (3) impacts were larger in places that facilitated work-directed activities, but that also administered real consequences on those who were neither gained economic independence nor accepted program assistance in working toward self-sufficiency.

[4] This study used a rigorous, but complex evaluation design, building on the optimal design modeling work done for the negative income tax evaluations (Watts & Rees, 1977).

[5] The Workforce Investment Act (PL 105-222) is now the major federally funded initiative to support workforce development services through statewide and local workforce investment systems (http://www.doleta.gov/usworkforce/wia/act.cfm).

The appropriate lens for framing questions and interpreting evidence can change. For example, early public policy interest in child care derived from a concern about work equity. An expanded supply of and financial support for child care was viewed as important in supporting the participation of women in the workforce. In this context, the emphasis was on policies that could expand the supply of safe, affordable, full-day care (Hayes, Palmer, & Zaslow, 1990). More recently, interest in child care stems from its potential as a means of reducing the school preparedness gap between children from poor and non-poor families. This places emphasis on the quality of child care, its academic content, and its availability to poor children, regardless of the employment status of their parents.

Some issues require multiple rounds of research using a variety of methods. For example, early efforts to address the economic and social consequences of teenage pregnancy and parenting were complicated by issues of basic values, by the influence of external agents such as parents and boyfriends, and by issues of normal adolescent development. It took no less than three multi-method demonstration program evaluations conducted over 12 sites to get a reasonable understanding of the problems associated with teen pregnancy and childbearing and the barriers to and opportunities for improving outcomes for teen mothers and their children, and to generate a collective body of evidence that supported major policy changes.

These three teenage parent demonstrations were conducted in a context where the social and economic demography of teenage parenthood was well understood. Ultimately, it was evidence from all three studies pieced together that guided the teenage parent provisions of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996.

The first of these studies tested a largely theory-driven model of intervention—providing services such as mentoring by caring adults, counseling, and referrals to educational services (Polit & White, 1988). It demonstrated little to no benefits for teen mothers.

The (not unreasonable) response to the null findings was to launch a second-generation demonstration testing an even richer array of services, including one-on-one mentoring, high-quality school-based child care, family planning services, and various education and social supports—a package of services that cost an average of over $4,000 a year per teenage mother (Quint, Polit, Bos, & Cave, 1994). Overall, this intervention failed to generate the intended benefits for the mothers or their children, and, for a subset of teen mothers, it had the unintended negative effect of hastening second births. However, it is notable that the impacts varied across sites in ways that prompted a careful examination of what might explain the pattern of program impacts across the test sites.

This supporting analysis uncovered a number of interesting findings—among them that a distinguishing feature of two programs that unwittingly hastened subsequent births was that they celebrated announcements of pregnancies and, more generally, tended to actively validate mothers' decisions, even when those decisions threatened longer-term prospects for their own health and well-being or that of their children. In contrast, the sites where pregnancy rates declined or were unaffected by the program exhibited more paternalistic behaviors, actively encouraging and empowering teen mothers to make decisions supportive of long-term goals—including delaying subsequent pregnancies and births.

The third generation demonstration grew largely out of public concern over the alarming rise in teen and out-of-wedlock childbearing, a growing public concern over high rates and costs of long-term welfare dependency, and a dearth of evidence

that public investment in self-sufficiency programs of various forms "works" (Ellwood, 1988; Ellwood, 1986; and Bane & Ellwood, 1994). This demonstration tested the consequences of mandatory school, job-training, or work for teenage mothers on welfare. The experimental findings showed very modest, favorable impacts of the intervention in some areas of interest and, importantly, no evidence of negative impacts for mothers or their children (Maynard, 1993). They also show some differential findings across subgroups that are correlated with the program climate in terms of enabling young mothers to continue their lives of dependence (for example, by lax enforcement of the activity requirements) versus taking a more paternalistic stance (for example, enforcing the requirements for engaging in self-sufficiency activities, actively working to alleviate barriers to employment, and administering consequences of noncompliance).

Collectively, these three demonstrations established the boundaries of influence for the teenage parent provisions incorporated in PRWORA—relatively modest benefits in terms of education and employment; modest delays in subsequent childbearing; and no adverse consequences for the young mothers or their children, such as higher poverty rates, problematic parenting, or worse developmental outcomes for children (Maynard, 1997; Kisker, Maynard, Rangarajan, & Boller, 1998; Granger & Cytron, 1999).

These studies also provided other important knowledge that was influential in changing attitudes and behaviors of the public, of policymakers, and of practitioners. The following are three examples. First, contrary to widely held perceptions, teenage mothers were generally willing and able to participate in school, job training, or employment. Commonly, self-confidence, expectations, and/or pressure from male partners contributed to nonparticipation in such activities. Second, access to contraceptives and contraceptive information is not sufficient for many young mothers to achieve personal (and public) goals to delay future childbearing. Erratic schedules, high rates of residential mobility, and pressure from male partners interfere with effective contraception. Third, attitudes and actions of public agency staff can promote and facilitate self-sufficiency-oriented behaviors of teen parents. But, for agency staff to serve these roles, it is important that their jobs be structured to explicitly make them responsible for providing such assistance and oversight of the outcomes for the teen parents.

Context matters in framing questions, designing research, and interpreting findings. This means that the applicability of evidence is bounded. Enthusiasm for grounding policy and practice in scientific evidence runs the danger of over-generalizing research results. A very poignant illustration is applying standards for the treatment of tuberculosis that are highly effective in the context of a developed nation to developing or undeveloped countries. Indeed, the highly effective treatment used in the United States could have quite devastating effects globally if applied in settings of high noncompliance (Kidder, 2003).

There are numerous other examples where over-generalizing the results of a study could be highly problematic. For example, the results of the experimental evaluations of abstinence education that are in process will contribute much-needed evidence to inform future policy and practice (Maynard et al., 2005). However, the immediate utility of the findings is limited by the fact that the evidence pertains to particular age groups of youth residing in communities characterized by particular norms, values, and services that likely affect in important ways how youth respond to the programs under study.

Some patterns of findings could lead to important definitive conclusions. For example, if the particular set of abstinence education programs being evaluated

proves successful in delaying sexual debut and in lowering sexual health risks for youth, this would certainly settle important questions about the *inevitability* that abstinence education *will* be harmful, as some have argued. It also would establish the potential for long-term benefits in terms of lower rates of teen pregnancies and births, fewer children born to teens and out-of-wedlock, and lower rates of sexually transmitted diseases. If the results are mixed (for example, delays in sexual debut, but higher rates of sexual health risks), if they are null (no impacts at all), or if they are consistently unfavorable, there will be a different set of lessons from the studies.

Whatever the pattern of results, the findings from these few studies currently underway will have limited generalizability. There are two reasons for this. First, all of the evidence will be from particular types of communities—for example, all are supportive of abstinence education and all volunteered to participate. Second, all of the evidence is from particular types of programs that, while varied in structure, content, and duration, deliver all or most of their services during the middle school years. Finally, the results need to be interpreted and used in the context of competing goals, scarce resources, and political realities. For example, increasing the intensity of services generally means serving fewer youth. Furthermore, some strategies, like condom distribution efforts, simply will not be permitted in some communities.

THE APPAM EDGE IN PRODUCING RELIABLE EVIDENCE

Evidence-based policy, management, and practice should draw on research that spans disciplinary and substantive boundaries—drawing, for example, on economics, political science, psychology, sociology, education, social welfare, communications, environment, and criminal justice. Moreover, it should use multiple forms of evidence. Standard evaluation texts tend to distinguish between intervention research (commonly referred to as quantitative) and implementation, process, and ethnographic research (commonly referred to as qualitative) (for example, see Rossi, Lipsey, & Freeman, 2004; Bloom, 2005; Orr, 1999; Light, Singer, & Willett, 1990; Krathwohl, 1998; and Mertens, 2005). These distinctions are not very relevant to most comprehensive evaluations that draw on multiple sources of data and methods of analysis. They also largely ignore the very important descriptive analysis that would, for example, provide a clear characterization of a public policy concern, in context.

Using multiple modes of inquiry to generate evidence on particular questions within a single study—solid descriptive framing data, reliable evidence of causal relationships, and rich information to contextualize the causal evidence and guide theoretical work—can improve the usefulness of findings and, in some cases, decrease the time and number of studies needed to generate useable evidence. When the full spectrum of information is not provided from a single study, end-users tend to piece together evidence from multiple sources even though the pieces may not really "fit" together. Alternatively, they may act on the basis of partial information, or delay acting at all. For example, much of the welfare reform research of the 1980s and early 1990s consisted of analyses based on available administrative data, which lacked important contextual information that could have been helpful, particularly in understanding the limits of program success and guiding "next-generation" efforts (Cottingham & Ellwood, 1989; Gueron & Pauly, 1991; Moffitt & Ver Pleog, 2001).

We know what constitutes good evidence and how to produce it. However, too often, decisions are made that compromise the credibility of the research findings.

Furthermore, even when studies are well designed to yield credible evidence on the right questions, reports sometimes miscast the findings in important ways. The most common example is presenting descriptive or correlational evidence in ways that invite interpreting it as evidence of causal relationships. Descriptive and correlational results can be powerful for identifying problems, characterizing and contextualizing issues, informing theory, guiding intervention planning, and monitoring the outcomes of programs and practices. They are not reliable sources of answers to questions of what would happen if a particular policy or practice were implemented.

Careful, descriptive analyses are extraordinarily important for policy and practice. Descriptive and correlational analysis can identify and characterizes economic, social, or educational problems in ways that focus policy attention on them and that guide the design of subsequent intervention research. Reliable descriptive statistics depend only on having valid measures for adequate size samples of relevant individuals. Useful correlational analysis can be generated through smart analysis of data that includes valid measures of the outcomes of interest for adequate size samples of individuals with particular characteristics, experiencing certain conditions, or living in particular contexts. In both cases, the analysis needs to use statistical procedures that account for the properties of the study sample and, when reported, findings should be accompanied by information regarding the reliability of estimates and the reference group to which they pertain.

Notable examples of landmark research of this genre include that by scholars such as James Coleman (1966), Frederick Mosteller and Daniel Patrick Moynihan (1972), and Christopher Jencks and colleagues (1972), who focused attention on the intergenerational nature of poverty and the role of educational disadvantage in that cycle. Their work stimulated decades of intervention research aimed at finding ways to compensate for educational disadvantages of children reared in low-resourced families.

Causal evidence answers the "so what?" question. It can be demanding to produce (Mosteller & Boruch, 2002), often engenders controversy (Ioanadis, 2005a, 2005b), and is quite susceptible to misuse. The high controversy and misuse quotients derive in no small part because of the closer link between intervention findings and policy recommendations. The principles for generating reliable causal evidence are simple and well known. The challenges come in being clear about the causal questions a study should address, in implementing the study as designed, and communicating the findings contextualized in a manner that minimizes errors in interpretation and application. The most successful studies are ones in which there is a real interest in knowing the answer to the causal questions being addressed and in which the study design and implementation are such that the evaluator is prepared to stand behind the findings regardless of what they are. Still, the smartest use of causal evidence generally involves some accumulation of findings across multiple studies and/or study settings. (See further discussion below of reasons for and methods of synthesizing findings across studies.)

In principle, generating reliable causal evidence requires valid measures of the outcomes of interest for individuals who differ only in their exposure to a particular condition—in the present case, a public policy or practice. There is broad agreement that, in many cases, it is quite feasible to achieve a close approximation to the necessary contrast through randomized controlled trials—studies in which relevant individuals or groups of individuals are assigned randomly to conditions representing the policy contrasts of interest. There also is general agreement on 10 key principles of rigorous randomized controlled trials:

1. Allocation to treatment and control conditions should be truly random (although it need not be 50/50).[6] Moreover, sample members should be retained in their originally assigned status throughout the study.
2. The details of the sample design and its implementation should be carefully documented and accounted for in the analysis. Researchers should maintain data on the assignment probability for all sample members and document blocking and/or clustering that was used in creating the study sample.
3. The minimum relevant impact should be determined a priori and the samples should be sufficiently large that there is a high probability that impacts of this size will be detected, should they occur.[7] Researchers should not invoke the "low power" excuse for null findings unless they are willing to invoke it equally for null findings that are consistent with and contrary to expectations.
4. Data sources should be the same for the contrast groups in the study. It is generally problematic to use information collected from different sources, for different reference periods, or using different settings or collection methods.
5. Every effort should be made to minimize sample loss over time. If sample attrition is other than trivial, it is important to examine and report evidence on the likelihood that sample loss might have biased the study results.
6. The analysis should be conducted in a manner that maximizes the reliability of impact estimates and minimizes the chance of bias in the findings. This means using multivariate analysis to measure impacts on the *full* sample the policy or practice was intended to affect.[8]
7. Confidence intervals around the impact estimates should account for statistical effects associated with the particular sample design used. This includes, for example, accounting for differential rates of assignment to intervention condition or differences between the units that were randomly assigned to intervention conditions and the units analyzed. An example would be accounting for "clustering effects" in the case in which classrooms are randomly allocated to the relevant contrast groups and students are the unit for the analysis (Bloom, 2005).
8. The results should be presented in easy-to-understand language and appropriately contextualized. At a minimum, this means presenting sample sizes, estimated means and standard deviations for the relevant contrast groups, differences in the means for the groups, and confidence intervals around those differences.
9. Results should be presented for important subgroups, if sample sizes are sufficient to support reliable conclusions. However, in interpreting results

[6] If one is tempted to rely on another method for constructing comparison groups, it would be wise to consider how comfortable it would be to defend "adverse" findings should they result. There is a large body of research demonstrating the unreliability of impact estimates generated using quasi-experimental study designs (for example, see Agodini & Dynarski, 2004; Glazerman, Mayer, & Myers, 2003; and Weisburd & Lum, 2001).

[7] A rule of thumb is that studies that randomize individuals need to have samples of about 500 in order to detect modest size impacts (about .2 standard deviations or 10 percentage points for a binary outcome with a mean of .5), and studies that randomize clusters generally should include at least 50 clusters. However, there are many factors that could raise or lower these numbers (see, for example, Bloom & Raudenbush, 2005; and Schochet, 2005).

[8] This is commonly referred to as "intention to treat" analysis. Those interested in measuring impacts for the "treated" group can, in many cases, generate reliable estimates from the full study sample findings following guidance provided in the econometrics literature. See, for example, Bloom (1984) and elaborations by Angrist, Imbens, and Rubin (1996).

for sample subgroups it is important to relate these findings to the overall study results.

10. Complementary nonexperimental analysis should be carefully distinguished from the causal findings. For example, results of a randomized controlled trial to examine the impacts of the offer of after-school services are much more credible evidence of the effects of the after-school programs than would be results of an exploratory sub-analysis looking at differential outcomes for those who take greater or lesser advantage of the after-school services.

The challenge, of course, is in implementing these various principles. Which of the many interesting questions will "drive" the study design? For example, should a teacher professional development study focus primarily on the consequences of the intervention for teacher practice or on the consequences of the intervention for student outcomes? Both are important. How close can you come to getting the optimal representation of programs/policies and participants in the study? Again, the evaluators likely will face some trade-offs. And, how close can you come to the ideal data collection strategy—optimal measures, information sources, and sample retention?

Disseminate findings to maximize usefulness and minimize controversy. The researcher generally has full control of the analysis and reporting of evidence, but not in its application. Still, the end-game for the researcher should be to produce the best *knowledge* to inform policy and practice, not the making of policy itself. It is the job of the policymakers and practitioners to use the knowledge produced to inform their work, which entails numerous considerations other that the scientific evidence on particular issues. Despite the best efforts on the part of researchers, evidence will be misused, as illustrated by the following two examples.

21st Century After-school Programs: This is a study designed to build a knowledge base over several years; yet, some attempted to apply early results to major policy decisions. This well-designed, large-scale program evaluation was intended to assess the potential benefits of offering youth in high-risk schools access to academically focused after-school programs designed and administered locally. It was not a test of a particular model of after-school program. Moreover, the study was designed as a multi-year evaluation, since any benefits could be expected to cumulate over time. The first report on program impacts showed no evidence of academic benefits of the programs over the first year following access to services (Dynarski et al., 2001).

Groups opposed to spending on after-school programs inappropriately used the first year study findings to support recommendations to cut federal support for after-school programs. The study was intended to evaluate the impacts of a policy as it played out in local practice, not the *potential* of a particular intervention. As such, many youth who were offered access to the after-school program did not participate—a fact that could change quite dramatically if parents had credible evidence that participation likely would improve their child's school performance. Furthermore, the results pertain to measured impacts after only one year of exposure to a program that may have been a start-up operation—factors that are noted in the report but ignored by the vested user groups.

In this case, the evidence was highly credible, thoroughly presented, and carefully contextualized. These facts blunted, somewhat, efforts to misuse the evidence, because both those in the scientific community and those policymakers who are interested in using evidence smartly could understand what the research did and did not reveal about the value of federal support for after-school programs.

Perry Preschool: Results of a study of a small-scale multi-faceted early intervention are over-generalized as evidence to support public funding for high-quality child care

or universal pre-K. This is a landmark study in two regards: (1) it was one of the first social experiments; and (2) it was one of the first tests of social policies aimed at addressing problems of economic, social, and cultural risks of young children (Schweinhart & Weikart, 1980; Barnett, 1991). This is a study of an intervention delivered to 58 youth in Ipsilanti, Michigan, during the late 1960s and 1970s. The study provides solid evidence of the *potential* of social and educational interventions to change the life-course trajectory of highly at-risk youth. It was not, however, a field test of child care, high-quality or otherwise, nor are the intervention results, *per se*, generalizable to conditions of today—facts that are more evident in some reports on the study than others.

There also are examples of studies that have been well designed to produce causal evidence, but where the outcome measures used were problematic; where the main analysis was conducted on a sub-sample defined by factors that are dependent on the intervention; and/or where some findings were never reported out because they were null or contrary to expectations. For example, much of the research on relationship interventions uses outcome measures that are heavily aligned with the interventions and the data collection strategies differ between the intervention and control groups (Reardon-Anderson, Stagner, Macomber, & Murray, 2004). Not infrequently, studies of pregnancy and sexually transmitted disease (STD) prevention programs report estimates of impacts on condom use for the subset of the study sample that is sexually active, ignoring the fact that most interventions aim to influence both patterns of sexual activity and contraceptive behavior among the sexually active (Scher et al., 2005). And, there is a whole literature pointing to evidence that studies yielding null or "contrary" findings, in particular, are less likely to be published than are those with statistically significant findings in the expected direction (Lipsey & Wilson, 1993; and Begg, 1994).

AGGREGATING EVIDENCE FOR POLICY PURPOSES

As the volume of evidence accumulates, there are both more opportunities for and greater challenges in sifting and sorting this evidence in meaningful ways to guide public policy. Formally aggregating evidence for policy purposes is common practice in medicine (see, for example, the Cochrane Collaboration library (http://www.cochrane.org/reviews/)). However, it is relatively new in other areas, such as education, social welfare, and criminal justice. Until the turn of this century research syntheses generally entailed one of two approaches: (1) inventory studies on a particular question with narrative summaries of study findings; or (2) chart the statistical results and pool them using meta-analytic methods to create weighted average estimates of causal impacts (Lipsey & Wilson, 2001; Cooper, 1998; and Cooper & Hedges, 1994).

Beginning at the turn of this century, there has been heightened interest in promoting evidence-based policies and practices. Essentially, the push is for a more formal approach to synthesizing the findings of numerous studies on a particular question than occurred in the evidence gathering that guided the policy decisions reflected in PRWORA, WIA, and No Child Left Behind (NCLB) (http://www.ed.gov/nclb/landing.jhtml?src=pb), for example. In the future, it is expected that major changes in policy directions will be guided by evidence, most often generated through a systematic review of the literature. Policies and practices will be decided, regardless of whether there is a lot or a little evidence to inform the process. However, if the available evidence has been smartly synthesized, decision makers will at least understand the extent to which they are operating in uncharted

territory, or a territory with equivocal, moderate, or strong support for the choices they are making.

An important question is: How do we sort, sift, and synthesize evidence to best inform those faced with important decisions of policy or practice? As with primary research, meaningful research syntheses may be largely descriptive, correlational, or focus on intervention effects. Indeed, my introduction to research syntheses was the work of Hanushek (1989, 1994, 1996, and 1997) and that of Hedges and his colleagues (1994a, 1994b). These two researchers reached different conclusions regarding whether money is important in determining academic outcomes of students. Only many years later do I understand that what appeared to be disputes over the interpretation of evidence was really differences in the questions each addressed.

As with any other type of analysis, it is important in conducting a research synthesis to focus on clearly defined questions. Then, the review on each question should include all evidence that meets specified standards for reliability and relevance. An example of such standards is those used by the What Works Clearinghouse (http://www.whatworks.ed.gov/). Once reliable evidence has been identified, it is important to contextualize this evidence in terms of its coverage of the populations and contexts of interest. For example, does all evidence on the effectiveness of a particular strategy of reading instruction for English language learners come from studies of native Spanish speakers or does it include a fair representation of speakers of other languages? Does the evidence pertain only to settings in which high levels of professional development were provided or does it include evidence based on varied levels of implementation and operational support? And, does it include only students in high performing districts or are the results based on students in a wide variety of settings?

As with primary intervention studies, syntheses of the results of such studies should recognize the greater credibility of causal evidence generated from well-implemented randomized controlled trials as contrasted with evidence from quasi-experiments or correlational research. The syntheses should pay careful attention to the nature of the research being examined and avoid pooling results across studies asking different questions or that pertain to quite different contexts. When evidence is pooled, it is important to understand and document the bounds of its applicability. And, the findings of syntheses should be reported out following guidelines similar to those noted above for the intervention studies themselves.

Two examples of where research syntheses have clearly added value are: (1) a synthesis of findings on the reliability of evidence from quasi-experimental studies versus randomized controlled trials (Glazerman et al., 2003); and (2) a synthesis of the effects of multi-systemic therapy (MST) (Littell, 2003). In the first instance, Glazerman and colleagues pursued evidence on two questions: (1) On balance, will well-designed quasi-experimental design studies yield evidence that is consistent with that from the more reliable randomized controlled trials; and (2) Is there any way to predict which quasi-experimental design studies will yield reliable evidence? The answer to the first question is yes and the answer to the second is no.

The importance of being both systematic and attentive to study quality when sifting, sorting, and synthesizing evidence is illustrated by the review of research on the effectiveness of MST. In contrast to the conclusion of most of the 23 prior reviews on the subject, Littell (2003) found that "MST may have been oversold. If MST has significant effects, they appear to be modest and unreliable" (page 29). Her experience led her to compile a list of "the 13 ways to draw wrong conclusions in a research review" (page 30). My eight favorites (paraphrased) are:

1. Limit the review to studies conducted by individuals with a vested interest in the results.
2. Ignore the fact that some sample has been inexplicably omitted from the analysis.
3. Ignore flaws in the implementation of random assignment and subsequent contamination of the intervention or control groups.
4. Ignore the fact that data for the intervention and control groups are not comparable.
5. Focus on immediate post program outcomes.
6. Include studies that have screened out noncompleters or no-shows.
7. Ignore high and/or differential sample attrition.
8. Restrict the review to studies reported in peer-reviewed journals or other easy to access sources.

It will be a long time until incontrovertible evidence exists to guide any major policy decision. Yet, the policy research community can contribute substantially to more effective policy and practice if we embrace the mission of advancing truth—including truth about what we do know and truth about when we are ignorant.

CONCLUSION

I want to conclude by reflecting back on the three personal experiences I shared to illustrate some of the challenges in using research to guide policy, management, and practice.

Do neighborhoods matter? Of course neighborhoods matter. The question is for whom and in what context. The proclamations of great influence come from questions that might guide targeting of resources or central planning. Poor neighborhoods are, indeed, homes to individuals with high rates of socio-economic disadvantage; they are places with relatively high rates of crime, and so forth. The findings of modest influence of neighborhoods come from analyses looking at the extent to which neighborhood resources compensate for other social or economic disadvantages or vice versa. Not surprisingly, the research suggests that some families are quite adept at overcoming the odds associated with living in poor neighborhoods, while other low-resourced families fail to thrive, even in supportive neighborhoods. Lastly, the conclusion that neighborhoods do not matter derives from evidence on the short-term effects of changes in neighborhood quality experimentally induced through giving families housing vouchers. This intervention stimulated a response by a modest proportion of families, it supported relatively modest improvements in neighborhood quality, and it did little to change other important environmental attributes, such as earnings and accumulated educational and social contexts. Each set of findings is important, but only when the evidence is applied appropriately.

Are pregnancy prevention programs effective? There is clear evidence that some programs are successful. However, none of the programs evaluated achieved a "homerun." Moreover, the vast majority of the studies have shown no evidence of effects. Notably, all of the estimated impacts are less that 15 percentage points in magnitude, and a nontrivial number of programs showed evidence of unintended "negative" consequences. The most important factors explaining the discordant conclusions from the various systematic reviews are the questions addressed by the research and the character of the evidence tha was counted. There is evidence that some programs are effective and we can even compile a list of the common charac-

teristics of those programs. However, the evidence suggests that the typical program studied was not effective, and that more studies than can be explained by chance show evidence of having adverse effects. If we were to compile the lists of common features for the "ineffective" and "harmful" programs, the lists would overlap with one another and with the list of features of the "effective" programs. Finally, by screening out significantly flawed studies, the evidence base shrinks quite substantially, and the estimate of the average impact is very small and not significantly different from zero.

So, should we give parents school vouchers or not? Whether school vouchers are good or bad is not going to be settled by any one, two, or three studies. One would expect impacts to vary depending on the value of the voucher; the levels and patterns of use; the private and public school climates; and family circumstances. It would be quite shocking if short-term academic performance effects would be sufficient in magnitude to definitively guide public policy. Rather, the New York City School Voucher Program evaluation and studies of voucher programs in several other cities offer complementary evidence on the benefits and limits of school choice and market forces in education. It will help the cause of evidence-based decision making if the research community distinguishes more clearly among those differences in findings that are due to the interventions themselves or the settings in which they were tested and those that relate to the untestable assumptions researchers invoked in their analysis. Commendably, the issues around the New York City School Voucher Program study were resolved in such a manner. Other disputes over evidence have been more confrontational. It is not surprising that the confrontational disputes tend to create distrust of research.

The APPAM conference is a terrific venue for intellectual exchange and debate among the methodologists, between methodologists and evaluators, and between the evaluators and the policymakers and practitioners. Evidence matters and it will matter more as we continue to raise the bar for what constitutes credible evidence, as more credible evidence accumulates, and as we become more facile and vigilant about synthesizing and disseminating research smartly.

REFERENCES

Agodini, R., & Dynarski, M.R. (2004). Are experiments the only option? A look at dropout prevention programs. The Review of Economics and Statistics, 86(1), 180–194.

Angrist, J.D., Imbens, G., & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434), 444–455.

Bane, M.J., & Ellwood, D.T. (1994). Welfare realities: From rhetoric to reform. Cambridge, MA: Harvard University Press.

Barnett, W.S. (1991). Benefits of compensatory preschool education. Journal of Human Resources, XXXVII(2), 279–312.

Begg, C.B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 399–409). New York, NY: Russell Sage Foundation.

Bloom, H.S. (1984). Accounting for no-shows in experimental designs. Evaluation Review, 8(2), 225–246.

Bloom, H.S. (2005). Randomizing groups to evaluate place-based programs. Learning more from social experiments: Evolving analytic approaches. New York: Russell Sage Foundation.

Bloom, H., & Raudenbush, S. (2005). Statistical analysis and optimal design for cluster randomized trials. New York: W.T. Grant Foundation.

Coleman, J.S. (1966). Equality of educational opportunity. Washington, DC: U.S. Department of Health Education and Welfare, Office of Education (U.S. Government Printing Office).

Cooper, H. (1998). Synthesizing Research (3rd ed., Vol. 2). Thousand Oaks, CA: Sage Publications, Inc.

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

Cottingham, P.H., & Ellwood, D.T. (Eds.). (1989). Welfare policy for the 1990s. Cambridge, MA: Harvard University Press.

DiCenso, A., Guyatt, G., Willan, A., & Griffith, L. (2002). Interventions to reduce unintended pregnancies among adolescents: Systematic review of randomised controlled trials. British Medical Journal, 324(7351), 1426–1430.

Dynarski, M., James-Burdumy, S., Mansfield, W., Mayer, D., Moore, M., Mullens, J., et al. (2001). A broader view: The national evaluation of the Twenty-First Community Learning Centers Program. Princeton, NJ: Mathematica Policy Research, Inc.

Ellwood, D. (1986). Targeting strategies for welfare recipients. Princeton, NJ: Mathematica Policy Research, Inc.

Ellwood, D. (1988). Poor support. New York: Basic Books.

Glazerman, S., Mayer, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Sciences, 589, 63–93.

Granger, R., & Cytron, R. (1999). Teenage parent programs: A synthesis of the long-term effects of the New Chance Demonstration, Ohio's Learning Earning and Parenting Program, and the Teenage Parent Demonstration. Evaluation Review, 23(2), 107–145.

Greenberg, D., & Shroder, M. (2004). The digest of social experiments (3rd ed.). Washington, DC: Urban Institute Press.

Gueron, J., & Pauly, E. (1991). From welfare to work. New York: Russell Sage Foundation.

Hayes, C.D., Palmer, J.L., & Zaslow, M.J. (Eds.). (1990). Who cares for America's children: Child care policy for the 1990s. Washington, DC: National Academy Press.

Hanushek, E.A. (1989). Impact of differential expenditures on school performance. Educational Researcher, 18, 45–51.

Hanushek, E.A. (1994). Money might matter somewhere: A response to Hedges, Laine, and Greenwald. Educational Researcher, 23(4), 5–8.

Hanushek, E.A. (1996). School resources and student performance. In G. Burtless (Ed.), Does money matter? The effect of school resources on student achievement (pp. 43–73). Washington, DC: Brookings Institution Press.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performace: An update. Educational Evaluation and Policy Analysis, 19(2), 141–164.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994a). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. Educational Researcher, 23(3), 5–14.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994b). Money might matter somewhere: A reply to Hanushek. Educational Researcher, 23(4), 9–10.

Hilsenrath, J.E. (October 24, 2005). Making waves: Novel way to assess school competition stirs academic row. Wall Street Journal, pp 1, 11.

Hollister, R., Kemper, P., & Maynard, R.A. (Eds.). (1984). The National Supported Work Demonstration. Madison, WI: University of Wisconsin Press.

Ioannidis, J. (2005). Contradicted and initially stronger effects in highly cited clinical research. Journal of the American Medical Assoication, 294(2), 218–228.

Ioannidis, J. (2005). Why most published research findings are false. PLoS Medicine, 2(8).

Jencks, C., Smith, C., Acland, H., Bane, M.J., Cohen, D., Gintis, H., et al. (1972). Inequality: A reassessment of the effect of family and schooling in America. New York: Basic Books, Inc. Publishers.

Kidder, T. (2003) Mountains beyond mountains: The quest of Dr. Paul Farmer, a man who would cure the world. New York, Random House.

Kirby, D.B. (2000). Emerging answers. Washington, DC: National Campaign to Prevent Teen Pregnancy.

Kisker, E.E., Maynard, R.A., Rangarajan, A., & Boller, K. Moving teenage parents into self sufficiency: Lessons from recent demonstrations. Princeton, NJ: Mathematica Policy Research, Inc., Document No. PR98-37.

Krathwohl, D.R. (1998). Methods of educational and social science research: An integrated approach (2nd ed.). Long Grove, IL: Waveland Press Inc.

Krueger, A.B., & Zhu, P. (2004). Another look at the New York City school voucher experiment. American Behavioral Scientist, 47(5), 658–698.

Light, R.J., Singer, J.D., & Willett,J.B. (1990). By design: Planning research on higher education. Cambridge, MA: Harvard University Press.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, education, and behavioral treatment: Confirmation from meta-analysis. American Psychologist, 48, 1181–1209.

Lipsey, M.W., & Wilson, D.B. (2001). Practical meta-analysis (Vol. 49). Thousand Oaks, CA: Sage Publications.

Littell, J.H. (2003). Systematic and nonsystematic reviews of research on the outcomes of multisystemic treatment. Paper presented at the 25th Annual APPAM Research Conference, Washington, DC.

Mayer, D.P., Peterson, P.E., Myers, D.E., Clark Tuttle, C., & Howell, W.G. (February 2002). School choice in New York City after three years: An evaluation of the school choice scholarships program. Final Report. Washington, DC: Mathematica Policy Research, Inc., Unpublished manuscript.

Maynard, R. (Ed.). (1993). Building self-sufficiency among welfare-dependent teenage parents. Report prepared for the U.S. Department of Health and Human Services Princeton, NJ: Mathematica Policy Research, Inc. http://aspe.hhs.gov/hsp/isp/tpd/folup1/synes1.htm.

Maynard, R. (1997). The role for paternalism in teen pregnancy prevention and teen parent services. In L. Mead. (Ed.), The new paternalism: Supervisory approaches to welfare (pp. 89–129). Washington, DC: The Brookings Institute Press.

Maynard, R., Ternholm, C., Johnson, A., Devaney, B., Clark, M., Homrighausen, J., & Kalay, E. (2005). First-year impacts of four Title V, Section 510 abstinence education programs.Unpublished manuscript, Princeton, NJ: Mathematica Policy Research, Inc. http://www.mathematica-mpr.com/publications/PDFs/firstyearabstinence.pdf.

Mead, L. (Ed.). (1997). The new paternalism: Supervisory approaches to welfare. Washington, DC: The Brookings Institute Press.

Mertens, D.M. (2005). Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods (2nd ed.). Thousand Oaks, CA: Sage Publications.

Moffitt, R.A., & Ver Ploeg, M. (Eds.). (2001). Evaluating welfare reform in an era of transition: Panel on data and methods for measuring the effects of changes in social welfare programs. Washington DC: National Academy Press.

Mosteller, F., & Boruch, R. (Eds.). (2002). Evidence matters: Randomized trials in education research. Washington, DC: The Brookings Institution.

Mosteller, F., & Moynihan, D.P. (1972). On equality of opportunity. New York: Random House.

Myers, D., Peterson, D., Mayer, D., Chou, J., & Howell, W.G. (2000). School choice in New

York City after two years: An evaluation of the School Choice Scholarships Program. Interim Report. Washington, DC.: Mathematica Policy Research, Inc.

O'Connor, A. (2002). Poverty knowledge: Social science, social policy, and the poor in twentieth-century U.S. history. Princeton, NJ: Princeton University Press.

Orr, L. (1999). Social experiments: Evaluating public programs with experimental methods. New York: Sage Publications.

Polit, D., & White, C. (1988). The lives of young, disadvantaged mothers: The five year follow-up of the Project Redirection Sample. Saratoga Springs, NY: Humanalysis, Inc.

Quint, J., Polit, D., Bos, H., & Cave, G. (1994). New chance: Interim findings on a comprehensive program for disadvantaged young mothers and their children. New York: Manpower Demonstration Research Corporation.

Reardon-Anderson, J., Stagner, M.W., Macomber, J.E., & Murray, J. (2004). A systematic review of the impact of marriage and relationship programs. Washington, DC: Urban Institute.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2004). Evaluation: A systematic approach (Seventh ed.). Thousand Oaks: Sage Publications.

Scher, L.B., Maynard, R.A., & Stagner, M.W. (2005). A systematic review of teen pregnancy prevention program evaluation findings. Philadelphia: University of Pennsylvania.

Schochet, P. (2005). Statistical power of random assignment evaluations of education programs. Princeton, NJ: Mathematica Policy Research, Inc.

Schweinhart, L.J., & Weikart, D.P. (1980). Young children grow up: The effects of the Perry Preschool Program on youths through age 15. Ipsilanti, MI: High Scope.

Turner, M.A., & Rawlings, L.A. (2005). Overcoming concentrated poverty and isolation. Washington, DC: Urban Institute.

Weisburd, D., Lum, C.M., & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? The Annals, 578, 50–70.

Watts, H.W., & Rees, A. (Eds.). (1977). The New Jersey income-maintenance experiment, Volume II; Labor-supply responses (Vol. II). New York: Academic Press.