Using Large-Scale Databases in Evaluation: Advances, Opportunities, and Challenges

William R. Penuel and Barbara Means

American Journal of Evaluation 2011 32: 118 originally published online 26 October 2010
DOI: 10.1177/1098214010388268

Article

American Journal of Evaluation
32(1) 118-133

DOI: 10.1177/1098214010388268
http://aje.sagepub.com

Using Large-Scale Databases in Evaluation: Advances, Opportunities, and Challenges

William R. Penuel[1] and Barbara Means[1]

Abstract

Major advances in the number, capabilities, and quality of state, national, and transnational databases have opened up new opportunities for evaluators. Both large-scale data sets collected for administrative purposes and those collected by other researchers can provide data for a variety of evaluation-related activities. These include (a) identifying or highlighting issues that invite greater client attention; (b) establishing the plausibility of policy theories of action; and (c) program evaluation. The authors illustrate through examples from the fields of education, social services, and public health both the opportunities provided by higher quality, interoperable data systems and the challenges encountered when using databases to identify issues of concern, test the plausibility of a proposed theory of action, or evaluate an existing program. The authors then explore implications for roles evaluators may need to play to support these uses of databases and for funding and infrastructure to support greater evaluator involvement in program planning and improvement.

Keywords

database, evaluation use, mixed-type evaluations, program improvement

The number, capabilities, and quality of state, national, and transnational databases have increased significantly over the past 25 years. Particularly relevant to practicing evaluators are the possibilities for accessing large-scale data sets collected by other researchers and linking multiple, interoperable databases. These capabilities allow evaluators to repurpose data sets to answer questions that the developers of databases did not anticipate but that may be important to clients. The improved quality of databases means that the data may be more trustworthy, and the linkages enable researchers to gain a more robust understanding about how different contexts can shape outcomes.

In this article, we focus on three kinds of evaluation uses that these databases have afforded evaluation researchers in recent years: identifying or highlighting issues that invite greater client attention; establishing the plausibility of policy theories of action and program evaluation. Evaluation and policy

[1]SRI International, Menlo Park, CA, USA

Corresponding Author:
William R. Penuel, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA
Email: William.penuel@sri.com

research audiences have employed extant databases for all three of these purposes, as the cases we present here show. For example, by linking data sets from different administrative sources, evaluation researchers have highlighted for community members the need for additional or targeted services to subpopulations whose outcomes are significantly worse than those of other groups. In addition, policy researchers have used large-scale databases to compare competing theories of action to guide policy regarding effective teaching practices. Finally, evaluators, working in close partnership with community members, have used databases as one source of evidence about the effectiveness of programs.

Before presenting our case examples, we review some of the key advances in the field that now permit evaluators to consider using databases in evaluation. The cases provide illustrations of potential ways for evaluators to take advantages of these advances for policy analysis, evaluation, and evidence-based decision making in conducting evaluation studies at different program stages, including prior to program development, and for both improvement and assessment functions, to draw on Chen's (1996) typology of evaluation types. The cases that we present highlight how researchers and evaluators have exploited the potential of databases but also illustrate some of their limits. Finally, in the last section we consider implications of the case examples for recasting the evaluator's role in projects that involve the use of large-scale databases.

Recent Advances in Large-Scale Databases

Three major advances in recent decades have made it possible for evaluators to consider their use in different aspects of program evaluation:

1. Many databases now collect longitudinal data on individuals, permitting evaluators to analyze change over time that may be linked to participation in programs.
2. Many databases now are interoperable; that is, they allow linking information from different data sources.
3. An infrastructure, including professional development, has emerged to support data use by both researchers and practitioners.

Below, we describe how each of these advances can help evaluators in planning and conducting evaluations of policies and programs.

The Potential Value of Databases With Longitudinal Data on Individuals

Longitudinal databases that include individual data on educational, developmental, or health outcomes permit evaluators to analyze change over time in outcomes, a key aspect of most evaluation studies. Clients typically ask evaluators to analyze whether outcomes of participants in programs improved and, when databases permit, to compare outcomes for nonparticipants to those of participants. In policy discussions, differences in outcomes for program participants and nonparticipants are often accepted as evidence that the program caused the observed differences. Claims about whether programs *caused* such improvements are not readily supported by longitudinal databases, however (Ginsburg & Rhett, 2003). Propensity score techniques (Rosenbaum & Rubin, 1983) for helping evaluators identify a matched comparison group, however, can increase the plausibility of claims that participation in a program was the primary cause of observed differences in desired outcomes. Most policy makers are comfortable making decisions based on this kind of evidence (Coburn, Honig, & Stein, 2009).

The Potential Value of Interoperable, Linked Databases

Scholars from across the diverse fields of medicine, public health, and education are increasingly recognizing that the different domains of individual well-being are deeply interconnected

(National Research Council and Institute of Medicine, 2002). Most administrative databases address some aspects of one of these domains but not all three. Given the complex interplay among the contributions of processes across different domains of well-being in producing important individual and social outcomes, databases that are interoperable and allow for linking of data on individual health, behavior, and academic attainment may provide richer resources for evaluators than longitudinal databases that include data from only one social sector or set of institutions (McLaughlin & O'Brien-Strain, 2008). In particular, linking databases may provide deeper insight into how family events and neighborhood characteristics may mediate the effectiveness of education or health programs (e.g., Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010). Linking databases may be particularly useful as well in highlighting issues of concern, including inequities in systems, which could lead to the identification of new programs (Guiton & Oakes, 1995).

The Potential Value of Processes to Support Database Use

Databases developed by individual teams of researchers or by governments are likely to be of little use if supports for their use are not also available. Supports for use include mechanisms by which evaluators and researchers can gain access to data and use them for multiple purposes. Such mechanisms are critical, since developers of databases cannot imagine ahead of time all of the purposes to which a database might be put. In addition, since the hope of many policy makers is that clinicians and practitioners, not just researchers and evaluators, will use databases to implement evidence-based practices (see, e.g., Roper & Tolleson-Rinehart, 2001), more specific supports may be necessary for end users, such as query tools on user-friendly web interfaces. End users may benefit from evaluator guidance in data interpretation (Johnson et al., 2009), including guidance in understanding the limits of databases for drawing inferences about the efficacy of particular programs or strategies.

Exemplars of Database Use to Support Key Functions of Evaluation

In this section, we present exemplars of database use that align to the three key functions that evaluations can play in providing data to inform policies and programs. These functions pertain not just to outcome evaluations that focus on judging the merit or worth of an evaluation but also to process and outcome evaluations at multiple phases of program development.

Many of the analyses of databases we describe below were not conducted by self-identified evaluators but by sociologists, public health researchers, and psychologists seeking to inform debates about policies and programs. In the final section, we take up the question of why exemplars of database use might be more readily found among these scholars than among evaluators and explore implications for the roles that evaluators may need to play in the future if they are to harness the potential of large-scale databases for evaluation.

Using Large-Scale Databases to Highlight Issues of Concern

Large-scale databases generally contain data on the indicators for which an organization is held accountable (e.g., student test scores in the case of school districts or patient survival rates in the case of surgical units). Typically, they also contain data elements on participant characteristics and on transactions that can be linked to outcome data in ways that highlight differential success rates. We present two illustrations of uses of databases to identify and highlight *issues of concern.* Issues of concern may pertain to difficulties that populations or subpopulations of persons may have in accessing services or they may pertain to outcome attainment differences. The two cases we present are both drawn from education but include use of data from sources outside education.

Montgomery County's Integrated Quality Management System and M-STAT Process: Identifying Inequities in Access to Advanced Coursework

Until fairly recently, the use of student data to make decisions at the district, school, and classroom levels was impeded by the lack of systems providing user-friendly access to student data (Wayman, 2005). In the last decade, however, several forces have converged to make data-informed decision making at all levels of the education system both more feasible and a priority. Improved data systems capable of tracking individual students' progress from year to year have become available and have been implemented in an increasing number of states and districts (Wayman, 2007). No Child Left Behind instituted requirements for achievement data reporting by student subgroup, requiring many districts to obtain or upgrade their student data systems, and has made schools and districts responsible for student achievement (Hess & Petrilli, 2006). The Institute of Education Sciences (IES) invested several hundred million dollars in grants to states to upgrade their student data systems in ways that would facilitate longitudinal analysis of student achievement data. Commercial entities responded to the need for new kinds of student information systems, data analysis tools, and data warehouses. As districts and schools have looked for strategies to help raise achievement, the use of data to predict and enhance student performance has emerged as perhaps the dominant improvement strategy. This emphasis on the use of student data did not fade with the change in Administrations after the 2008 election. The Obama Administration has increased the emphasis on the use of education databases, making the use of student data to improve instruction one of the central components of its education strategy and inviting states and districts to compete for federal funds to support this activity through the Race to the Top and Investing in Innovation programs (Duncan, 2009).

National surveys at the district and school levels have confirmed the rapid growth in educators' access to student data systems: Large-scale national surveys found that the proportion of teachers reporting that they had access to an electronic student data system rose from 48% to 72% between 2006 and 2007 (Gallagher, Means, & Padilla, 2008). By 2009, 74% of teachers reported having remote access to student data and 62% reported having used this access (Gray, Thomas, & Lewis, 2010). But surveys also reveal shortcomings in the kinds of data available to school staff. Most commonly, teachers had access to state achievement test data for the students they had the prior year; only 11% could access data for their current set of students in 2007 (Gallagher et al., 2008). A national survey of district staff found that nearly all of them had electronic student information systems (providing real-time access to student attendance and schedule information), 93% had an electronic system containing students' state assessment data and 72% had systems containing data on student performance on district tests. But less than half of the districts surveyed in 2008 reported having systems that would allow them to link student outcome data to process data (such as particular interventions experienced by a student), and only 23% had systems that would enable linking student outcomes to financial expenditures (Means, Padilla, & Gallagher, 2010).

The ability to link outcome data to other data is critical to making the most out of data systems, as the experience of Montgomery County Public Schools (MCPS) illustrates. Located just outside Washington, District of Columbia, MCPS historically served a population that was mostly White and affluent. The district enjoyed a reputation as one of the best school systems in the country. As new immigrant populations moved into Montgomery County in the 1980s and 1990s, though, increasing numbers of students lived in poverty and came from immigrant families or historically underserved ethnic groups. In 1999, when Jerry Weast was appointed as the new MCPS superintendent, the difference between the performance of African American and Hispanic students on one hand and White and Asian students on the other was as much as 45% on key measures such as passing seventh-grade math (Thompson & Winking, 2007). Weast was given the mandate to address the achievement gap.

Studying the district in detail—both by the numbers and by riding to every district mail stop in the county—Weast discovered what he later characterized as a "Red Zone" district within the district. The Red Zone consisted mainly of immigrant families, Hispanics, and African Americans living in urban pockets of poverty centered on major transportation arteries. Looking at data for schools within the Red Zone versus that for the rest of the district's schools, Weast found a 20-point difference in average scores on standardized tests, a difference in the proportion of students enrolled in honors and Advanced Placement courses, and differences in college enrollment (Childress, Doyle, & Thomas, 2009). The differences were stark in grade 3 reading scores and got no smaller as students progressed through the system.

In 2000, Weast appointed a chief information officer for the district with the goal of providing teachers and administrators with the data and the resources they would need to address the achievement gap. Work began on an Integrated Quality Management System (IQMS) that would combine student information such as enrollment, attendance, grades, scheduling and test performance with data on professional development, finances, and human resources. An Instructional Management System (IMS) was added to the IQMS so that teachers had online access to curriculum resources, student performance on formative and summative assessments and various screenings, along with curriculum guides and lesson plans.

In addition to the data warehouse, the district invested in tools to help teachers diagnose and respond to their students' learning needs. A joint venture with a technology developer led to software for a handheld computing device that allowed teachers to do individual assessments of a student's literacy skills and capture the data for the student's electronic file. Having the data available electronically, reading coaches could look at data for individual students, identify students who were not making the expected progress, and proactively seek out the relevant teachers to offer suggestions and support. Since initiating this intervention, the district has experienced reading achievement gains for all its students but especially large gains for African American, Hispanic, and low-income students (Childress et al., 2009).

More recently, the district has implemented a collaborative, data-focused process that it calls M-STAT. In one application of this process, MCPS community superintendents came together to examine Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) participation and advanced course enrollment rates within their high schools. The district had long recognized the lower participation of Hispanic and African American students in advanced courses and had switched from using teacher recommendations to using PSAT/NMSQT scores and other objective indicators to counsel students into honors and Advanced Placement (AP) courses. The M-STAT process revealed that these student groups were less likely than other students to participate in the PSAT/NMSQT examination, though, and the schools did not have any strategies in place to encourage their participation. The community superintendents then started working with their principals to institute steps such as meeting with African American and Hispanic parents to talk about the importance of the PSAT/NMSQT and to making the examination experience more enjoyable by providing snacks before and afterward. The district's deputy superintendent and one of the community superintendents developed the Honors/AP Potential Identification Tool that uses measures such as grades, PSAT/NMSQT, and scores on other standardized tests to identify minority students with good prospects for success in advanced courses. The deputy superintendent identified students who appeared qualified for advanced courses but were not enrolled in them and started interviewing individual students to understand the barriers to enrollment. Progress is being made: in 2008, the school district reported that 88% of African American and 84% of Hispanic students in MCPS took the PSAT/NMSQT; over 60% of African American and Hispanic high school students were enrolled in at least one honors or AP course (Childress et al., 2009).

As this example illustrates, the availability of a data system is a necessary but not sufficient condition for improvement. The longitudinal data enabled the district leadership to show not only

that there was a big achievement gap but also that the gap was persistent. Linking the data on achievement to geography and to course enrollment data, furthermore, the district was able to pinpoint where there were opportunities to improve opportunities for students of color to take advanced classes that could help close the gaps. But it was the superintendent's leadership in focusing on equity issues and the M-STAT process that were most important in determining the kinds of data to examine and developing action plans based on data.

The Youth Data Archive (YDA): Highlighting Differential Outcomes for Foster Youth

The YDA is an initiative aimed at helping communities improve youth development outcomes by linking data on individual youth over time and across different institutional settings. Partners in the YDA include school districts; community college districts; county agencies including education, human services; city agencies including recreation and parks departments; and youth-serving nonprofit organizations. The John W. Gardner Center for Youth and Their Communities (JGC) at Stanford University, in collaboration with the SPHERE Institute, houses the archive, develops agreements with nonprofit and government agencies in selected counties and communities in Northern California, and facilitates groups' investigation of available data. Some of these data are reported and available in statewide databases, but at the community level, the YDA can incorporate additional data that agencies are not required to report but that they collect to monitor and improve services (London & Gurantz, 2010). In addition to linking administrative data on youth outcomes, in some communities the JGC staff develop and collect data on setting characteristics, such as the motivational climate of schools and afterschool programs, which can be linked to these data and provide information that can be useful to organizations seeking to understand the relationship between program participation and youth outcomes (McLaughlin & O'Brien-Strain, 2008).

As in the MCPS example, the YDA example is first and foremost one of a process for supporting continuous improvement of systems for promoting youth development rather than simply a repository of data. The JGC works in collaboration with local agencies to articulate a set of research questions of concern to the community and to identify sources of data that could help address those questions. Staff develop memoranda of understanding that detail what data are to be included, what analyses are to be performed, and how and with whom analyses will be shared. All agreements comply with laws regarding the protection of privacy and human subjects. The JGC works closely with its partners to seek external funding to support activities related to the YDA.

The kinds of expertise required of JGC staff go beyond simply being able to articulate research and evaluation questions, collecting data, and performing appropriate analyses of results. To be successful, staff must be able to work collaboratively with partners to identify answerable questions that are both of interest to community members and of potential interest to the broader field of youth development. They must be skilled in negotiating data sharing agreements related to highly sensitive data with partners who may not know the JGC and who may be suspicious of some of the other agencies that are partners in the work. They must also be able to sustain partnerships' focus through the process of data analysis, interpretation, and identification of next steps for improving the system of opportunities in youth development in a community. Along the way, different agencies express varied and changing levels of commitment to the process of improvement, requiring staff to adjust strategy, analysis, and communication plans to remain attuned and aligned to the goals of partners.

The payoff for this multifaceted work can be high, especially when the YDA permits analyses of youth outcomes across time and setting that answer important policy questions. For example, JGC staff recently collaborated with several agencies in San Mateo County (California) to analyze educational outcomes for court-dependent youth in foster care (Castrechini, 2009). The frequent school

and residence changes typical of this group make tracking outcomes difficult without a tool like the YDA; for this particular analysis, JGC staff were able to link dependency records from Child Welfare Services to educational data from several school districts. The analysis produced, as might be expected, a picture showing that outcomes for court-dependent youth were much worse than those for other children, an analysis made possible only by linking two longitudinal data from two different administrative data sets. At the same time, the detailed records of the YDA provided some insight into the ways that children's placements were related to outcomes. In general, outcomes were better for youth placed in involuntary family settings than for those placed in out-of-home settings (Castrechini, 2009). As a consequence of the analysis, JGC staff produced and conversation they facilitated about the findings, the collaborating agencies came to view the need for academic support for these youth, especially in group homes and other nonfamily settings, as requiring more attention and as an urgent need to address in the community.

Using Large-Scale Databases to Test the Plausibility of Policy Theories of Action

Prior to conducting evaluations of policies and programs, evaluators sometimes conduct *evaluability assessments* (Leviton, Khan, Rog, Dawkins, & Cotton, 2010; Wholey, 2004). One goal of evaluability assessment is to establish that program goals and theories of action are plausible, given the realities of programs and the settings in which they are implemented. Oftentimes, evaluators conduct such assessments prior to conducting evaluations or implementing policies, relying principally on available sources of data, which could include large-scale databases. Typically, a proposed policy is based on the assumption that instituting practice A will cause desired outcome B. Although analyses of extant data cannot prove that A causes B, such analyses can be used to test whether there is a relation between A and B. If not, and all other relevant factors are held constant, then there may be a flaw in the policy theory. Correlation between A and B that would be expected if A does in fact cause B does in face exist in the data, other factors being held constant. In the examples below, we present one use of longitudinal databases of student achievement to argue for the plausibility of a particular set of strategies for improving teacher quality, a plausibility argument that has influenced recent policy making in education. In the second example, we present two policy analyses intended to inform debates about the efficacy of the policy.

Using Student Achievement Databases to Analyze Teacher Quality

A core goal of the Obama Administration's education policy is the improvement of teacher quality. The Administration's definition of teacher quality centers on teachers' effectiveness: a high-quality teacher is one whose students make strong achievement gains on standardized tests (Robelen, 2009). This definition stands in contrast to the previous Administration's definition of teacher quality, which emphasized strong subject matter preparation (Hess & Petrilli, 2006), and with definitions of teacher quality defined in terms of the quality of a teacher's instructional practice (e.g., Kennedy, 2005). Researchers' arguments about the importance of measuring and selecting teachers on the basis of test score growth, developed in large part through their analyses of longitudinal databases of student achievement data, have figured strongly in the change in policy, as is evident in the remarks below by Secretary of Education Arne Duncan on the importance of linking teacher and student data:

> Hopefully, some day, we can track children from preschool to high school and from high school to college and college to career. We must track high growth children in classrooms to their great teachers and great teachers to their schools of education.
>
> In California, they have 300,000 teachers. If you took the top 10 percent, they have 30,000 of the best teachers in the world. If you took the bottom 10 percent, they have 30,000 teachers that should probably

> find another profession, yet no one in California can tell you which teacher is in which category. Something is wrong with that picture. (Duncan, 2009)

Behind Duncan's claims are multiple analyses of state databases of achievement, constructed over many years, that focus on the so-called value added of teachers to achievement. Tennessee was the first state to experiment with linking teacher and student data in an accountability system (Sanders & Horn, 1994), and the architect of its system, William Sanders, was one of the first researchers to explore patterns in a student database to estimate individual teachers' effectiveness. In one study (Sanders & Horn, 1998), Sanders and his team concluded that the best predictor of student gain was not student background but rather what teacher a student had. Moreover, the team found that teacher effects were both additive and cumulative, with little evidence that an effective teacher encountered later in a student's career could offset effects of ineffective teachers encountered at an earlier age.

As more states implemented longitudinal data systems in the early 2000s, more analysts began looking for similar patterns in achievement data, and their findings were similar to those of Sanders' team. For example, Rivkin, Hanushek, and Kain (2005) examined achievement data from three cohorts of elementary and middle school students in Texas in the 1990s and found large variation among teachers with respect to gain scores of their students. These gains were not associated with a teacher's experience (beyond the first 3 years of teaching) or education nor could they be attributed entirely to student background characteristics. A study of student achievement in mathematics in an urban district in Texas (Hanushek, Kain, O'Brien, & Rivkin, 2005) similarly found wide variation in teacher effectiveness. Those researchers found that a student who has a teacher at the 85th percentile of effectiveness, as estimated by their model, could expect annual gains roughly one fifth of a standard deviation greater than those of a student with an average teacher (i.e., one who is at the median in terms of student gains).

The researchers who conducted these analyses have advocated for policies focused on selecting, deselecting, and rewarding teachers on the basis of ''value added.'' Hanushek (2009), in particular, has argued for basing pay on the achievement gains made by teachers' students and for delaying teachers' tenure until their value added can be established from analyzing student test score growth. The analyses have caught the attention of education reformers (as well as assessment researchers critical of the analyses and concerned about policy applications of the approaches [Baker et al., 2010]), and they have become part of the rationale for the Obama Administration's policies regarding accountability and teacher pay.

The National Longitudinal Study of Adolescent Health

The National Longitudinal Study of Adolescent Health, or Add Health, is a study of adolescents who were in grades 7 to 12 during the 1994–1995 school year. There have been four waves of data collection on the cohort's social, economic, psychological, and physical health and academic attainment. In addition, the Add Health team has connected data on a variety of family, neighborhood, school, friendship, peer group, and romantic relationship factors to investigate how these shape development into young adulthood. The study team, managed by researchers at the University of North Carolina Population Center, collected the data both through in-school questionnaires and in-home interviews.

The power of the database derives from its sample and the fact that the team has followed students across many years. The original sample was selected as a stratified, random sample of 80 high schools from across the United States. The study team also recruited for the study one feeder middle school, that is, a school that sent graduates to the high school. Thus, the sample is a nationally representative sample that reflects the socioeconomic and cultural diversity of the United States in the

mid-1990s. In addition, the database includes data from four waves of data collection, the most recent wave collected in 2008. Some data from the first three waves are available for public use. The longitudinal data collected through Add Health allow investigators to pose questions that few evaluators charged with analyzing short-term impacts of programs ever have the data to answer.

For example, investigators have used Add Health data to analyze the efficacy of a strategy for sexuality education popular in the mid-1990s called "the abstinence pledge." The Southern Baptist Convention initiated the pledge movement in 1993 with its "True Love Waits" campaign (for the Convention's chronology of this campaign, see http://www.lifeway.com/tlw/history.asp). Though the pledge was not a formal program, many sexuality education programs endorsed the idea of an abstinence pledge as a key component of what would become "abstinence-only" models of sexuality education. These models taught young people that abstinence was the only guaranteed means to protecting sexual health and did not teach about contraception. The salience of the pledge was heightened by the fact that in 1996, the U.S. government required grantees of two major federally funded programs to adopt abstinence-only models of sexuality education. Though it is impossible to separate the pledge movement or federal requirements from the moral and political controversies that shaped them (Fine & McClelland, 2006; Santelli et al., 2006), researchers with access to Add Health data have sought to use data to shed light on these controversies.

Like evaluators who conduct evaluability assessments to compare program goals with program realities, sociologists sought to use data from Add Health to examine where the program theory behind abstinence-only "collides with reality," as they put it (Brückner & Bearman, 2005; for a similar kind of analysis of Add Health data examining the theory behind mentoring programs, see DuBois & Silverthorn, 2005). In an initial analysis of Add Health data, two researchers explored whether or not taking an abstinence pledge delayed the initiation of sexual intercourse among adolescence (Bearman & Brückner, 2001). They found that adolescents who took an abstinence pledge were more likely to delay first intercourse for several months; on average, the pledge was associated with a lower baseline rate of time to sexual initiation by 34%. At the same time, pledging worked only in early and middle adolescence, they found, and when there were neither too few nor too many pledgers in the adolescent's school. Further, those who broke their pledges were at greater risk for sexually transmitted diseases, since they did not use contraception at as high a rate as nonpledgers when they did engage in sexual intercourse. On the basis of their analysis, the researchers concluded:

> Critics of the pledge movement are as concerned about adolescent sex as are the supporters of the movement. They are both wrong and right about the pledge. They are wrong when they think that it does not work. But they are right when they think that it cannot work as a universal strategy . . . . The contextual effect we identify suggests interesting limits to the applicability of a universal, pledge-based policy. Like most other things, pledging works in moderation. (Bearman & Brückner, 2001, p. 902)

During a follow-up survey in 2001–2002, respondents provided urine samples, and researchers analyzed whether adolescents who had taken a pledge had similar rates of sexually transmitted diseases as nonpledgers (Brückner & Bearman, 2005). The researchers had predicted that pledgers would have lower rates of disease, since they would have had fewer different partners. Contrary to their expectations, however, they found no difference between the two groups, even though pledgers had initiated sexual intercourse later, had less cumulative exposure, fewer partners, and fewer nonmonogamous partners on average. The reason, they found, was that although pledging delayed first intercourse, most (88%) pledgers went on to engage in premarital sexual intercourse, and they were less likely to use contraceptives when they did so. Further, the pledgers were less likely than nonpledgers to be tested for sexually transmitted diseases, a risk factor for further sexual health difficulties in the future.

Efficacy trials conducted and reviewed subsequent to these analyses would provide less supportive evidence for abstinence-only programs. One review of randomized controlled trials found mixed evidence from three different studies for a delay in initiation of sexual intercourse stemming from abstinence-only program participation (Bennett & Assefi, 2005). Another set of analyses of a different longitudinal database, the National Survey of Family Growth, reported similar findings to the Add Health analyses regarding sexually transmitted diseases: participation in abstinence-based sexuality programs did not reduce adolescents' risk, when compared to the risk of adolescents who participated in more comprehensive programs that included information about contraception (Kohler, Manhart, & Lafferty, 2008). Characterizing U.S. policy on sexuality education in light of this evidence, Constantine (2008) described the country as being "left behind" other countries, whose policies were more evidence-based.

Using Large-Scale Databases to Evaluate Programs

Evaluating the efficacy of policies and programs is a key function of evaluation. A well-designed efficacy study can provide evaluators with evidence for making judgments about the merit or worth of a program, evidence that policy makers can use to make decisions about the fate of programs (Dynarski, 2008). At the same time, efficacy trials, especially those that employ random assignment designs, are expensive and not always feasible to implement even when resources are available to do so. Under such conditions, interoperable databases with longitudinal data on individuals and their participation in programs can be a useful tool for analyzing evidence about the efficacy of programs.

An important qualifier regarding the potential of large-scale databases pertains specifically to their use for program evaluation. Rarely do databases include indicators that are closely aligned to goals of specific programs, and in practice, many evaluators have found it necessary to develop and support complementary, local indicators to measure program inputs and success. In developing and supporting these indicators in such a way that they become part of the larger data infrastructure, evaluators who use databases successfully find themselves in a different position with respect to their roles as evaluators. They find, for example, that they serve as "data intermediaries" who are responsible for securing agreements among different institutions for data sharing and as full partners with school systems, community groups, or organizations driven by a commitment to improvement that create an appetite for the kind of data they are able to develop through database linkages. Conversely, where no such commitment exists, stakeholders may take a defensive posture toward data, avoiding interpretations of results that might call into question program designs or implementation strategies.

In the example below, we describe one study conducted by researchers at the University of Chicago in partnership with the Chicago Public Schools to use databases in conjunction with complementary indicators of program success collected by the researchers to support evaluation of programs in that school district.

The Consortium of Chicago School Research's Teacher Induction Study

The Consortium on Chicago School Research is a research center housed at the University of Chicago and is a partnership among researchers and staff of the Chicago Public Schools. It was formed in 1990, and its initial focus was on examining the effects of a district-wide implementation of decentralized governance in Chicago. Then, as now, its focus remains on Chicago Public Schools, though its studies reach a wide audience of policy makers and researchers outside the district. A core component of the Consortium's mission is to build strong ties between research and practice, by bringing evidence to bear on concerns of policy makers and leaders in the district (Bryk et al., 2010). Its activities include the design of indicators to track progress toward policy and program

objectives, investigations of whether and how policies and programs are working, and building capacity for evidence-based decision making through engagement of a wide range of researchers, policy makers, and practitioners in the process.

Like the John W. Gardner Center, the Consortium maintains a comprehensive data archive to support research. The archive integrates administrative records from the school district as well as data the Consortium collects on a regular basis from schools. One key source of data in the archive is a biannual survey of teachers, which asks teachers to respond to questions about their school context and experience of teaching. These data serve as an important complement to district demographic data on schools and individual student achievement data. Having data on context, teaching, and learning outcomes allows the Consortium to develop analyses that can investigate how policies aimed at changing the context (e.g., organizational reforms) of schools affects teaching and student achievement (Bryk et al., 2010). In addition, as in the example presented below, analyses of contextual data can support investigations of the conditions under which programs targeting teachers can be effective in supporting their work.

A study conducted in 2007 on teacher induction illustrates one way that the Consortium uses its data archive to support program evaluation goals (Kapadia, Coca, & Easton, 2007). The district had recently instituted a comprehensive induction program, the GOLDEN Teachers program, which targeted all novice teachers in the district. With funding from the Joyce Foundation, the Consortium collaborated with district researchers to examine the influence of induction program participation on three outcomes: Novices' teaching experience, intent to continue in the profession, and plans to remain teaching in school. The key sources of data for the study were district administrative records and the biannual survey of teachers that the Consortium administered. The simple answer to the question of whether induction program participation was related to outcomes was "yes." For example, 47% of novice elementary school teachers participating in an induction program intended to continue teaching in their school, compared to 39% of novices not in a program, a statistically significant difference. But when researchers included contextual variables in their model, they found that simply participating in induction programs was insufficient to affect teachers' plans to stay in their school. Teachers who reported weak leadership and classrooms with a high percentage of students with behavior problems were less likely to plan to stay in their school, as were teachers in schools where administrative records showed there were high concentrations of children living in poverty. These problems largely offset the potential benefits of mentoring programs for novice teachers in the district.

Several aspects are noteworthy with respect to how the Consortium used its data archive to inform Chicago Public Schools about how well the induction programs in which its teachers were participating were working. First, the likelihood that the district would use the results was high, given that the analysis focused on an issue of great concern to the district; the report indicates that in Illinois, nearly 40% of teachers leave the profession within the first 5 years of their careers, costing school districts significant resources to replace them. Second, the complementary sources of data available to the researchers (multiple years' worth of administrative records and researchers' own biannual survey of teachers) allowed them to analyze key links in a policy theory of action, namely, whether teachers in induction programs experienced them as providing key components past research had found to be important and whether the effects of mentoring programs were conditioned by contextual variables that might diminish their effects.

Future Directions and Challenges Ahead for Evaluators in Using Large-Scale Databases

We have highlighted a number of ways that databases can support evaluation in the examples presented above. Our examples serve both improvement and assessment functions, and many of the

researchers who conducted the analyses would not likely identify themselves as evaluators. We explore why the examples that we considered the most compelling shared these characteristics, with an eye toward identifying new roles and tasks for evaluators that are implied by the potential of databases for evaluation.

Why Not Just Outcome Evaluations Focused on Assessing Program Worth?

In addition to outcome-assessment evaluations, we have identified examples of database use that highlight issues of concern for policy makers to inform program design and that provide data on the plausibility of program and policy theories of change after policies and programs have been implemented. Analyses of databases that serve these latter two purposes are likely to be valuable either before a program is designed or in the earliest stages of design. We conjecture that the main reason why such uses are of great potential value to evaluators in the long term is that they harness the unique potential afforded by longitudinal and interoperable databases and may be useful ways for policy makers to gain an understanding of program needs and implementation, when resources needed for a separate efficacy or effectiveness trial are not available.

The Consortium on Chicago School Research example is a case of what Chen (1996) would call a ''mixed-type'' evaluation in which the aim was to improve both implementation and outcomes for induction programs in the district. The study was not designed as an evaluation of the GOLDEN Teachers program per se, and the fact that all teachers were targeted and some participated in multiple programs would have made this a difficult goal to accomplish. Rather, its aim was to understand whether participation was related to teachers' plans to stay in the profession and their schools and to explore relationships between desired outcomes and key features of those programs. In this respect, their analyses showing that the quality and perceived helpfulness of different induction program activities mattered for teachers' plans are consistent with larger, nationwide survey results showing a strong relationship between teacher retention and access to helpful mentoring supports (Smith & Ingersoll, 2004).

Experimental studies of impact remain the ''gold standard'' in the fields of medicine, public health, and education for determining the efficacy of programs. On their own, longitudinal databases do not permit evaluators to rule out the possibility that some unmeasured confounding variable is the cause of a change in desired program outcomes. Many databases do not include information on who participated in which programs, and many do not include data on specific short-term outcomes that individual programs seek to impact. These limitations make large-scale databases far less useful for summative evaluation than for informing policy and program design, where identification of need and establishing the plausibility of a particular design are more critical steps. These latter two functions can be well supported by large-scale databases, as the examples above show, and policy makers and program designers may be convinced by evidence from analyses of large-scale databases when no efficacy studies have yet been conducted on a particular type of program design or to address a particular need.

Why Not Evaluators?

In the examples above, sociologists, public health researchers, and psychologists all figured prominently in analyses of databases, while evaluators, by contrast, played roles mainly in the Chapin Hall example, which focused on summative evaluation. The reasons why, we conjecture, are several. First, evaluators often find that clients approach them only when programs are well developed and need a summative evaluation. By contrast, policy researchers in each of the disciplines named above often seek to influence program design over the long term, by looking at patterns of relationships in applied settings that are implied by theories and evidence from their respective disciplines. Second, most evaluators address questions that are appropriate to programs

and settings at a much smaller scale than the kinds of questions that can be asked using data from large-scale databases. The outcomes of interest to evaluators are typically shorter term, and participants in programs may not always be included in databases. In the Consortium on Chicago School Research evaluation example presented above, the databases relied on district-wide data, but evaluators supplemented data from personnel databases with their own primary data collection.

The Way Forward: New Roles and Tasks for Evaluators

Evaluators can and should play significant roles in orchestrating and supporting large-scale database use in evaluation. Most evaluators relish the opportunity to become involved in program design. Moreover, providing program planners with insights derived from social science research can make the theories of action guiding programs more plausible and robust (Donaldson, 2007). Evaluators can analyze the outcome data available in databases to assess how well aligned they are to specific program and policy goals. Relying on such sources, when they are available, can greatly improve the efficiency of evaluation studies and reduce their cost to clients.

Making the most of large-scale databases implies new roles for evaluators, however. One such role is that of a ''data intermediary,'' that is, someone who helps secure data agreements among different agencies so as to be able to link different data sources that may be relevant to an evaluation. Securing such agreements requires great political skill and patience, as those researchers and evaluators who have sought to develop them have discovered. Another role is that of a long-term partner to policy makers and program developers. Long-term partnerships enable evaluators to get in on the ground floor of programs, to influence their design in ways that might be informed by large-scale databases. A third is ''facilitator of sense-making'' to community members who are likely to need help posing questions of databases and understanding what analyzes can and cannot tell them about participant needs and the efficacy of particular programs and policies.

Additional tasks will be required of evaluators to make the most of large-scale databases as well. To fulfill the role of data intermediary, evaluators will need to identify potential data sources from different sectors, identify the people and institutions that can grant them access to the data, and secure agreements from all parties to use the data for their intended purposes. As long-term partners, evaluators may need to be involved to a much larger degree than is typical in helping programs identify relevant research and conduct evaluability assessments using data from large-scale databases. They may need to undertake iterative cycles of measurement development in partnership with program staff as well, to ensure that evaluations include measures that are closely aligned to program goals. Finally, evaluators may need to develop participatory processes for sense making, so that community members can make sense of data together. These roles are more likely to emerge when program stakeholders develop or share a deliberative notion of how evaluation can inform the democratic process, that is, through honest, but value-driven discourse among free and equal participants (Hanberger, 2001; Lehtonen, 2006).

Many evaluators already play such roles and engage in such tasks, and theory-based evaluation (Donaldson, 2007), participatory evaluation (Cousins & Earl, 1992), and empowerment evaluation theory (Fetterman, 2001) imply that these roles and tasks are critical. But limited funding for evaluation and the short-term nature of grants to programs for their operation make the kinds of long-term partnerships that facilitate the ongoing use of databases for evaluation difficult to achieve. Needed are new kinds of infrastructures that enable more enduring partnerships that bridge the worlds of research and practice (Donovan, Wigdor, & Snow, 2003), as well as policy environments that encourage the use of evaluation for program improvement, not just accountability. If such conditions can be met, we see a bright future for the use of large-scale databases in evaluation.

Declaration of Conflicting Interests
The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding
The author(s) received no financial support for the research and/or authorship of this article.

References
Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E. H., Ladd, H. F., Linn, R. L., ... Shepard, Lorrie, A. (2010). Problems with the use of student test scores to evaluate teachers. EPI Briefing Paper #278. Washington, DC: Economic Policy Institute.

Bearman, P., & Brückner, H. (2001). Promising the future: Virginity pledges and first intercourse. *American Journal of Sociology*, *106*, 859-912.

Bennett, S. E., & Assefi, N. P. (2005). School-based teenage pregnancy prevention programs: A systematic review of randomized controlled trials. *Journal of Adolescent Health*, *36*, 72-81.

Brückner, H., & Bearman, P. (2005). After the promise: The STD consequences of adolescent virginity pledges. *Journal of Adolescent Health*, *36*, 271-278.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago.* Chicago, IL: University of Chicago Press.

Castrechini, S. (2009). *Educational outcomes for court-dependent youth in San Mateo County.* Stanford, CA: John W. Gardner Center for Youth and Their Communities.

Chen, H. T. (1996). A comprehensive typology for program evaluation. *American Journal of Evaluation*, *17*, 121-130.

Childress, S. M., Doyle, D. P., & Thomas, D. A. (2009). *Leading for equity: The pursuit of excellence in Montgomery County Public Schools.* Cambridge, MA: Harvard Education Press.

Coburn, C. E., Honig, M. I., & Stein, M. K. (2009). What's the evidence on districts' use of evidence? In J. D. Bransford, D. J. Stipek, N. J. Vye, L. M. Gomez, & D. Lam (Eds.), *The role of research in educational improvement* (pp. 67-87). Cambridge, MA: Harvard Education Press.

Constantine, N. A. (2008). Converging evidence leaves policy behind: Sex education in the United States. *Journal of Adolescent Health*, *42*, 324-326.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*, 397-414.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications.* Mahwah, NJ: Erlbaum.

Donovan, S., Wigdor, A. K., & Snow, C. E. (2003). *Strategic education research partnership.* Washington, DC: National Research Council.

DuBois, D. L., & Silverthorn, N. (2005). Natural mentoring relationships and adolescent health: Evidence from a national study. *American Journal of Public Health*, *95*, 518-524.

Duncan, A. (2009). *Robust data gives us the roadmap to reform.* Paper presented at the Fourth Annual Institute of Education Sciences Research Conference.

Dynarski, M. (2008). Bringing answers to educators: Guiding principles for research syntheses. *Educational Researcher*, *37*, 27-29.

Fetterman, D. M. (2001). *Foundations of empowerment evaluation.* Thousand Oaks, CA: Sage.

Fine, M., & McClelland, S. I. (2006). Sexuality education and desire: Still missing after all these years. *Harvard Educational Review*, *76*, 297-338.

Gallagher, L. P., Means, B., & Padilla, C. (2008). *Teachers' use of student data systems to improve instruction: 2005 to 2007.* Washington, DC: Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, U.S. Department of Education.

Ginsburg, A. L., & Rhett, N. (2003). Building a better body of evidence: New opportunities to strengthen evaluation utilization. *American Journal of Evaluation*, *24*, 489-498.

Gray, L., Thomas, N., & Lewis, L. (2010). *Teachers' use of educational technology in U.S. public schools: 2009 (NCES 2010-040)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis*, *17*, 323-336.

Hanberger, A. (2001). Policy and program evaluation, civil society, and democracy. *American Journal of Evaluation*, *22*, 211-228.

Hanushek, E. A. (2009). Teacher deselection. In D. Goldhaber & J. Hannaway (Eds.), *Creating a new teaching profession* (pp. 165-180). Washington, DC: Urban Institute.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality.* NBER Working Paper 11154. Cambridge, MA: National Bureau of Economic Research.

Hess, F. M., & Petrilli, M. J. (2006). *No child left behind.* New York, NY: Peter Lang.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*, 377-410.

Kapadia, K., Coca, V., & Easton, J. Q. (2007). *Keeping new teachers: A first look at the influences of induction in the Chicago Public Schools*. Chicago, IL: Consortium on Chicago School Research.

Kennedy, M. M. (2005). From teacher qualifications to teaching quality. *Educational Leadership*, *63*, 14-19.

Kohler, P. K., Manhart, L. E., & Lafferty, W. E. (2008). Abstinence-only and comprehensive sex education and the initiation of sexuality of teen pregnancy. *Journal of Adolescent Health*, *42*, 344-351.

Lehtonen, M. (2006). Deliberative democracy, participation, and OECD peer reviews of environmental policies. *American Journal of Evaluation*, *27*, 185-200.

Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N., & Cotton, D. (2010). Evaluability assessment to improve public health policies, programs, and practices. *Annual Review of Public Health*, *31*, 213-233.

London, R. A., & Gurantz, O. (2010). *State and local data infrastructure for tracking secondary to postsecondary educational outcomes.* Stanford, CA: John W. Gardner Center for Youth and Their Communities.

McLaughlin, M. W., & O'Brien-Strain, M. (2008). The youth data archive: Integrating data to assess social settings in a societal sector framework. In M. Shinn & H. Yoshikawa (Eds.), *Toward positive youth development: Transforming schools and community programs* (pp. 313-332). New York, NY: Oxford University Press.

Means, B., Padilla, C., & Gallagher, L. P. (2010). *Use of education data at the local level: From accountability to instructional improvement.* Washington, DC: Office of Planning, Evaluation, and Policy Development, U.S. Department of Education.

National Research Council and Institute of Medicine. (2002). *Community programs to promote youth development.* Washington, DC: National Academy Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*, 417-458.

Robelen, E. W. (2009, April). Obama echoes Bush on education ideas. *Education Week*, *28*, 118–19..

Roper, W. L., & Tolleson-Rinehart, S. (2001). Health care data and health: From numbers to outcomes. *Pharmacoepidemiology and Drug Safety*, *10*, 363-366.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299-311.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*, 247-256.

Santelli, J., Ott, M. A., Lyon, M., Rogers, J., Summers, D., & Schiefler, R. (2006). Abstinence and abstinence-only education: A review of policies and programs. *Journal of Adolescent Health*, *38*, 72-81.

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, *41*(3), 681-714.

Thompson, S., & Winking, D. (2007). Raising the bar, closing the gap. *Strategies*, *13*, 3-16.

Wayman, J. C. (2005). Involving teachers in data-driven decision making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed at Risk*, *10*, 295-308.

Wayman, J. C. (2007). Student data systems for school improvement: The state of the field *TCEA Educational Technology Research Symposium: Volume 1* (pp. 156–162). Lancaster, PA: ProActive.

Wholey, J. S. (2004). Evaluability assessment. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 33-62). San Francisco, CA: Jossey-Bass.