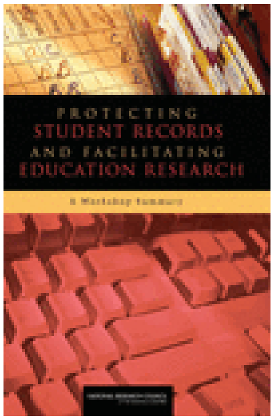

Protecting Student Records and Facilitating Education Research: A Workshop Summary

ISBN: 0-309-12800-5, 0 pages, , ()

This free PDF was downloaded from:
http://www.nap.edu/catalog/12514.html

Visit the National Academies Press online, the authoritative source for all books from the National Academy of Sciences, the National Academy of Engineering, the Institute of Medicine, and the National Research Council:

- Download hundreds of free books in PDF
- Read thousands of books online, free
- Sign up to be notified when new books are published
- Purchase printed books
- Purchase PDFs
- Explore with our innovative research tools

Thank you for downloading this free PDF. If you have comments, questions or just want more information about the books published by the National Academies Press, you may contact our customer service department toll-free at 888-624-8373, visit us online, or send an email to comments@nap.edu.

This free book plus thousands more books are available at http://www.nap.edu.

Copyright © National Academy of Sciences. Permission is granted for this material to be shared for noncommercial, educational purposes, provided that this notice appears on the reproduced materials, the Web address of the online, full authoritative version is retained, and copies are not altered. To disseminate otherwise or to republish requires written permission from the National Academies Press.

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

PROTECTING STUDENT RECORDS AND FACILITATING EDUCATION RESEARCH

A Workshop Summary

Margaret Hilton, *Rapporteur*

Committee on National Statistics

Center for Education

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Award No. 20071157 between the National Academy of Sciences and the Ewing Marion Kauffman Foundation, Award No. 9010 between the National Academy of Sciences and the William T. Grant Foundation, and Award No. 200800057 between the National Academy of Sciences and the Spencer Foundation. The work of the Committee on National Statistics is supported by a consortium of federal agencies through a grant from the National Science Foundation (Number SBR-0453930). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-12799-8
International Standard Book Number-10: 0-309-12799-8

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Research Council (2009). *Protecting Student Records and Facilitating Education Research: A Workshop Summary.* Margaret Hilton, Rapporteur. Committee on National Statistics and Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

PLANNING COMMITTEE ON PROTECTING STUDENT RECORDS AND FACILITATING EDUCATION RESEARCH

FELICE J. LEVINE *(Chair),* American Educational Research Association, Washington, DC
ROBERT F. BORUCH, The Wharton School, University of Pennsylvania
HELEN F. LADD, Terry Sanford Institute of Public Policy, Duke University
MARTIN ORLAND, WestEd, Washington, DC
JAY PFEIFFER, Florida Department of Education, Tallahassee
BARBARA SCHNEIDER, College of Education, Michigan State University

THOMAS J. PLEWES, *Study Director*
MARGARET HILTON, *Senior Program Officer,* Center for Education
MICHAEL J. SIRI, *Senior Project Assistant*
PAULA SKEDSVOLD, *Director of Education Research Policy,* American Educational Research Association (until August 2008)

COMMITTEE ON NATIONAL STATISTICS
2007-2008

WILLIAM F. EDDY (*Chair*), Department of Statistics, Carnegie Mellon University
KATHARINE ABRAHAM, Department of Economics and Joint Program in Survey Methodology, University of Maryland
WILLIAM DUMOUCHEL, Lincoln Technologies, Inc., Waltham, Massachusetts
JOHN HALTIWANGER, Department of Economics, University of Maryland
V. JOSEPH HOTZ, Department of Economics, Duke University
KAREN KAFADAR, Department of Statistics, Indiana University
DOUGLAS MASSEY, Department of Sociology, Princeton University
SALLY MORTON, RTI International, Research Triangle Park, North Carolina
VIJAY NAIR, Department of Statistics and Department of Industrial and Operations Engineering, University of Michigan
JOSEPH NEWHOUSE, Division of Health Policy Research and Education, Harvard University
SAMUEL H. PRESTON, Population Studies Center, University of Pennsylvania
KENNETH PREWITT, School of International and Public Affairs, Columbia University
LOUISE RYAN, Department of Biostatistics, Harvard University
ROGER TOURANGEAU, Joint Program in Survey Methodology, University of Maryland, and Survey Research Center, University of Michigan
ALAN ZASLAVSKY, Department of Health Care Policy, Harvard University Medical School

CONSTANCE F. CITRO, *Director*

Acknowledgments

This report is a summary of a workshop on protecting student records and facilitating education research convened by the National Research Council (NRC) and the American Educational Research Association. The workshop would not have become a reality without the generous support of the Ewing Marion Kauffman Foundation, the Spencer Foundation, and the William T. Grant Foundation.

We thank our colleagues who served on the planning committee, each of whom brought deep and varied expertise to the process of planning the workshop. Their diverse expertise on protecting student records and facilitating education research added greatly to the success of the endeavor. Although the planning committee played an important role in designing the workshop, they did not participate in writing this report.

We are especially grateful to the experts who quickly responded to our request for background papers on research and privacy issues, including Gerald Gates, U.S. Bureau of the Census (retired); Jane Hannaway, Urban Institute; Helen Ladd, Duke University, Susanna Loeb, Stanford University; Clara Muschkin, Duke University; and Steven Winnick, Nelson Mullins Riley & Scarborough. Their papers provided a substantive context for the rich discussions that took place at the workshop.

We also thank the many experts, including members of the workshop planning committee, who participated as presenters and panelists: Thomas R. Bailey, Columbia University; Ellen Campbell, U.S. Department of Education; Lizanne DeStefano, University of Illinois at Urbana-Champaign; Myron Gutmann, University of Michigan; Helen Ladd, Duke

University; Roberta Ness, University of Pittsburgh; Kenneth Prewitt, Columbia University; Ross Santy, U.S. Department of Education; Barbara Schneider, Michigan State University; Marilyn Seastrom, U.S. Department of Education; Jeff Sellers, Florida Department of Education; and Jonathan White, Agency for Healthcare Research and Quality.

This workshop summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process. We thank the following individuals for their review of this report: Eric Hanushek, Hoover Institution, Stanford University; Stephen Plank, Department of Sociology and Baltimore Education Research Consortium, Johns Hopkins University; Barbara Schneider, College of Education, Michigan State University; and Steven Winnick, Nelson Mullins Riley & Scarborough, Washington, DC.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft of the report before its release. The review of this report was overseen by Eleanor Singer, Institute for Social Research, University of Michigan. Appointed by the National Research Council, she was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author and the institution.

We are grateful for the leadership and support of Michael Feuer, executive director, and Miron Straf, deputy director of the NRC's Division of Behavioral and Social Sciences and Education (DBASSE), and Constance Citro, director of the Committee on National Statistics. We also thank Paula Skedsvold, former director of education research policy at the American Educational Research Association, Margaret Hilton, senior program officer in the NRC Center for Education, Christine McShane, senior editor, and Michael Siri, senior project assistant in the Committee on National Statistics.

Felice J. Levine, *Chair*
Thomas J. Plewes, *Study Director*
Planning Committee on Protecting Student
Records and Facilitating Education Research

Contents

Appendixes

1

Introduction

In August 1974, President Gerald R. Ford signed the Family Educational Rights and Privacy Act (FERPA) into law. One of two privacy laws Congress approved that year in response to the breach of public trust created by the Watergate scandal, FERPA was designed to protect the privacy of individual student test scores, grades, and other education records (U.S. Code, Title 20, Chapter 31, Section 1232g).[1]

Much has changed since that time. Education policies now emphasize education standards and testing to measure progress toward those standards, as well as rigorous education research. At the same time, private firms and public agencies, including schools, have replaced most paper records with electronic data systems. Reflecting the movement toward electronic data, many social science researchers have changed their methods; today, they may conduct fewer original surveys to gather research data and turn more often to administrative data maintained in government databases.

These trends have converged to greatly increase the supply of data on student performance in public schools. With funding from the U.S. Department of Education, the states are compiling student records from local schools and districts into statewide databases, with unique student identifiers that can be used to track students' performance as they move through grade levels and schools. Although these databases represent a

---

[1] Text of the act is available at http://www4.law.cornell.edu/uscode/html/uscode20/usc_sec_20_00001232---g000-.html.

rich source of longitudinal data, researchers' access to the individually identifiable data they contain, as well as to student record data maintained at the local level by individual schools and school districts, is limited by the privacy protections of FERPA. Researchers' limited access to individual student data slows research not only in education but also in related fields, such as child welfare and health.

To explore possibilities for data access and confidentiality in compliance with FERPA and with the Common Rule for the Protection of Human Subjects, the National Academies and the American Educational Research Association convened the Workshop on Protecting Student Records and Facilitating Education Research in April 2008 (see Appendix A for the workshop agenda). The workshop was supported by the Ewing Marion Kauffman Foundation, the Spencer Foundation, and the William T. Grant Foundation.

To carry out the workshop, the National Academies' Committee on National Statistics and Center for Education appointed an expert planning committee chaired by Felice J. Levine, researcher and executive director of the American Educational Research Association. The planning committee was charged to

> Plan for a workshop at the National Academies on providing research access to administrative records (including test scores) pertaining to elementary, secondary, and higher education students and their schools while protecting the privacy and confidentiality of the information. The planning committee will be charged with commissioning papers for presentation, and convening and serving as moderators for the workshop.

WORKSHOP GOALS AND FRAMEWORK

Felice Levine opened the workshop by welcoming all participants and providing an overview of the key issues to be discussed. Over the past five years, researchers have become increasingly interested in accessing the state education databases that compile student records, particularly because the No Child Left Behind Act of 2001 requires that "scientifically based [education] research" drive state and local use of federal education funds. These concerns informed the central question of the workshop—how to reconcile FERPA protections with current educational needs and goals. Levine explained that the workshop would address this central question in a broader context, examining approaches to reconciling privacy protections with research access not only in education, but also in other fields, such as health care.

Levine observed that the workshop was timely, because the Department of Education was seeking comments on proposed changes to its FERPA regulations. The proposed new rules address not only when schools and colleges can release student information for the purpose of protecting health and safety (following the April 2007 massacre at Virginia Polytechnic Institute and State University), but also when student information may be released for research purposes (U.S. Department of Education, 2008a). Levine invited thoughtful discussions that would inform useful comments on the proposed new rules. She mentioned that comments on the rules by the American Educational Research Association[2] would reflect an online survey of its members about FERPA, which drew a large response from over 250 education researchers. In closing, she predicted that the workshop would be valuable, observing that the National Academies' Committee on National Statistics has a long history of successfully addressing issues of research access and privacy protection (e.g., National Research Council, 1993).

ACCESS AND PRIVACY IN CONTEXT

Miron Straf (National Research Council) described the larger context of data privacy and research access issues surrounding the workshop. His remarks reflected a series of reports issued by the Committee on National Statistics over the past three decades. Straf began by defining the following key terms (Bradburn and Straf, 2003):

Information: knowledge, facts, or representations of them.
Personal information: information that is or can be linked directly or indirectly to some person. Identifiable information is personal information.
Data: information that is collected, compiled, captured, created, or received for one or more purposes.
Confidential data: data with personal information.
Statistical data: data without any personal information.
Disclosure: the release of personal information.
Discovery: to become aware of personal information from statistical data and other knowledge.
Privacy: an individual's control over who has access to information about him or her. The concept of privacy is relevant to what personal information becomes data.

---

[2] Joint comments on the proposed rule, submitted by the American Educational Research Association, the American Statistical Association, and the Consortium of Social Science Associations, were published following the workshop (American Educational Research Association, 2008).

Confidentiality: protection against the release of personal information. An important distinction is that privacy pertains to individuals; confidentiality to their information.

Straf then distinguished between (1) confidential data, with personal information, and (2) statistical data, without any personal information. He said that privacy pertains to the boundary between personal information and data and to the release of confidential data to others. For example, an individual may be willing to provide personal information to a health provider but opposed to having that same personal information provided to his or her employer.

On the basis of this analysis, Straf argued that, although people have the right to control their personal information, they do not have the right to control statistical data derived from that information. For example, the fact that a parent has a child enrolled in the eighth grade of a particular school district is personal information, but the number of eighth graders reported by the school district is an example of statistical data. An individual parent has no right to exclude his or her child from that count. Expanding on this analysis, Straf argued that it is not a violation of confidentiality to produce statistical data from one's personal information and, more broadly, it is not a violation of privacy or confidentiality to use statistical data for a purpose different from the one for collecting the information from which the statistical data were derived.

However, two key problems remain, according to Straf. The first is disclosure of personal information in confidential data, and the second is discovery of personal information from statistical data when those data are combined with other knowledge. He outlined two approaches to protect confidential data against both problems:

1. Altering the data in one of several ways, such as removing personal identifiers, collapsing individual data categories, adding random errors (statistical noise) to the data, or creating replicated (synthetic) data.
2. Restricting access—one approach is to license researchers who are then subject to penalties for disclosure or discovery. Straf noted that the National Center for Education Statistics has been a leader in using this approach (see Chapter 4). Another approach is to provide access at highly protected sites (data enclaves, research data centers), where analyses and other outputs are screened before they are released to any researcher.

Straf explained that new variants of restricted access have emerged. In one, at the request of a researcher, agency staff members analyze confidential data and then screen the results before releasing them to the researcher online. In another, Cornell University economist John Abowd has created a virtual research data center, which provides access to synthetic data over the Internet (Cornell University, 2008). The National Opinion Research Center at the University of Chicago has created a virtual data enclave that licenses researchers for restricted access online.

Although these protective approaches are important, Straf said, they also impose new costs and risks. Protected research sites, such as the Census Bureau's research data centers, are not easily available to many researchers, and, even when they are, they may not provide access to all relevant data. In addition, researchers are unclear about the extent to which replicated data corresponds to real data.

Straf then gave a brief overview of federal privacy laws and regulations. The Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) is designed to protect confidentiality of data collected for statistical purposes by government agencies. In addition, 17 federal agencies have adopted the Federal Policy for the Protection of Human Subjects, widely referred to as the Common Rule (U.S. Code of Federal Regulations, Chapter 45, Section 46). The Common Rule requires universities, federal agencies, and other research organizations to establish institutional review boards (IRBs). These boards review proposals to conduct research involving human participants, and they may reject proposals or require alterations in order to ensure adequate protection of individual privacy and data confidentiality. Straf said that, although these boards sometimes impose unnecessary confidentiality requirements that delay valuable research activities, they are important, providing an extra level of protection against disclosure and discovery of personal information. Most boards provide expedited reviews for research proposals not seen as involving significant risk.

Straf said that IRBs often require researchers to ensure that written, informed consent will be obtained from individuals who provide personal information for use in research studies. Informed consent documents are designed to clarify who will have access to the personal information and how it will be used. Straf argued that informed consent should apply only to personal information and should not be required when a researcher wants to use statistical data derived from that information (because, in his view, privacy and confidentiality do not apply to statistical data). Nevertheless, informed consent is valuable to build trust with the public, he said. Straf suggested that informed consent documents clearly describe all potential uses of the data sought, including research uses (National Research Council, 1993), and refer to the larger goals of the research, such as to improve the quality of education.

Arguing that statistical agencies' goal of zero tolerance for disclosure of confidential data is unrealistic, Straf suggested that agencies might instead adopt a standard of reasonable care, which would balance a small risk of disclosure against the great benefit of social science research. An additional protection would result if the agency placed the onus on the researcher using the data to avoid discovery or disclosure of personally identifiable information, as the National Center for Education Statistics does in its licensing arrangements (see Chapter 4).

Straf said that current tensions result from education policy makers' "voracious" demand for data on student performance and the Department of Education's efforts to promote rigorous research. Advances in cognitive, behavioral, and neurosciences are opening up new research avenues with potential to reduce disparities in educational achievement, he said, but researchers need access to education records to pursue these avenues. Describing many state education agencies as "wary" of providing education records for research because of uncertainty about how FERPA applies, he observed that many different methods are available to provide research access while protecting personal information, which is the goal of FERPA.

DISCUSSION OF KEY ISSUES

In response to Straf's presentation, Levine said that the key issue is the migration of individual data to statistical data. As discussed in *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data* (National Research Council, 2007), when researchers link statistical data sets to other data sets (such as geospatial data), they sometimes create personally identifiable data without having the consent of the individual whose data are now identifiable. When conducting a small-scale survey, a researcher routinely obtains each survey participant's informed consent for the uses of the data (including a warning about possible disclosure or discovery), but when there is no consent process, as is the case with administrative data, it is unclear how to allow research access while protecting privacy.

Straf's proposal for a reasonable standard of care led to a discussion of breaches of confidential data. Robert Boruch (University of Pennsylvania) said that the United States Privacy Protection Study Commission created under the Privacy Act of 1974 had searched for disclosures or risks of disclosure and found them only in marketing surveys and other private-sector information-gathering activities. Gerald Gates (Census Bureau–retired) said that, although the Census Bureau does not document disclosures, it has a staff dedicated to studying data files in order to determine whether links to external data would reveal individual identities, and this staff has identified some dangers. Myron Gutmann (Inter-

University Consortium for Political and Social Research) said that a recent search for lawsuits related to data disclosure had uncovered very few (National Research Council, 2007). As the director of a data-archiving organization, Gutmann said, he had asked other data-archiving organizations for examples of disclosures but found none. When Gutmann directed his staff to study the consortium's data files, the staff found few dangers. Gutmann explained that, because survey data has "noise," the odds of disclosure are small and that some papers on this topic would soon be published.

Barbara Schneider (Michigan State University) agreed that it is hard to personally identify data in large national data sets, but said state-level data are "easier to crack," raising critical confidentiality issues. This led to discussion of whether state databases of deidentified education records might actually include personally identifiable information, as well as the risks of disclosure from these databases.

REPORT OVERVIEW

This report continues in Chapter 2 with discussion of the Department of Education's current interpretation of FERPA and proposed new regulations to carry out the law, along with a description of the department's initiative to assemble and report state educational performance data. Chapter 3 discusses the value of education research using student and school records using examples of three studies that promise to inform needed improvements in public schooling. Chapter 4 presents models that allow researchers to access education records in ways that protect confidentiality and discusses the limitation of these models. Building on that discussion, Chapter 5 describes similar models of research access and confidentiality protection in other sectors, discussing as well the limitations of these models. The final chapter includes reflections about key issues and next steps by members of the workshop planning committee and other workshop participants.

2

Balancing Privacy, Confidentiality, and Access at the U.S. Department of Education

This chapter illuminates tensions between the privacy and confidentiality goals of the Family Educational Rights and Privacy Act (FERPA) and the goal, included in the No Child Left Behind Act, of using education data for research and accountability. The first section describes the Department of Education's current approach to implementing FERPA as it affects data access for research, as well as its proposed new FERPA rules. The second section discusses a Department of Education initiative to assemble and publicly report state educational achievement data and the challenges of doing so while protecting privacy and confidentiality. The final section sketches a successful approach to providing research access while protecting privacy and confidentiality, used in the department's own National Center for Education Statistics.

CURRENT INTERPRETATION AND PROPOSED NEW REGULATIONS

Ellen Campbell (U.S. Department of Education) provided an overview of the law and regulations (U.S. Code, Title 20, Chapter 31, Section 1232g). FERPA directs schools and higher education institutions to protect the rights of parents and students (age 18 or entering college) to inspect and review education records, to seek to amend education records, and to consent to disclosure of personally identifiable information from education records. The law applies to education agencies that receive funding from the department, including public elementary and secondary schools, some private elementary and secondary schools, and public or private institutions of higher education. FERPA provisions apply to a wide variety of education records, including medical records maintained by school health professionals. Provisions of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L. 104-191), as well as FERPA regulations, clarify that school medical records protected by FERPA are not governed by HIPAA.

FERPA defines "personally identifiable information" to include not only the student's name and the name of the student's parent or other family member and address of the student or student's family, but also a personal identifier, such as a social security number or student number or a list of personal characteristics or other information that would make the student's identity easily traceable. Although one of the primary rights of parents and students protected by FERPA is to consent to disclosure of personally identifiable information, such consent is not required under certain exceptions. Two of these exceptions are sometimes applied in decisions granting access to education records for research purposes:

- Disclosure to federal, state, and local educational authorities conducting an audit, evaluation, or enforcement of education programs (U.S. Code, Title 20, Chapter 31, Section 1232g, Subsection b).
- Disclosure to "organizations conducting studies for, or on behalf of, educational agencies or institutions for the purpose of developing, validating, or administering predictive tests, administering student aid programs, and improving instruction" (Ibid).

The law requires a school or higher education institution to maintain a record of each request for access to and each disclosure of an education record. In addition, when disclosing information from education records, the school should inform the receiving party that the information may not be further disclosed (with some exceptions).

Campbell said that the Department of Education expects that its proposed new FERPA regulations (U.S. Department of Education, 2008a) will improve access to education data for research and accountability purposes. The new rules would make it easier for state and local education agencies to redisclose information to each other, such as when a state department of education discloses student K-12 education records to a state higher education commission in order to track individual student achievement over time. The proposed rules would also update and clarify the definition of personally identifiable information and provide standards for removing all personally identifiable information from education data, as necessary or appropriate to release the information as deidentified data. A state education agency's release of properly deidentified data (meeting these standards) to an outside researcher would not constitute a disclosure requiring prior informed consent. In addition, the regulations would permit a state education database or education agency to attach a code to properly deidentified information that allows an education researcher to match information received from the same source. Finally, the proposed rule provides new information and recommendations for safeguarding records contained in electronic data systems.

Steven Winnick (Nelson Mullins Riley & Scarborough) argued that, despite some improvements in the proposed rules, FERPA continues to present significant barriers to the use of state data systems to improve public education. He noted that both federal and state policies call for using these data systems to advance standards-based education reform and that the department's Institute of Education Sciences had awarded $50 million to the states in fiscal year 2008 to develop and maintain these data systems.

Winnick explained that he works closely with the Data Quality Campaign, a national, collaborative effort to encourage states to compile longitudinal databases of education information (National Center for Educational Achievement, 2008). Its goal is not to water down the protections of FERPA, but to facilitate use of education records for research and accountability purposes. The major challenge to advancing this goal, in his view, is that FERPA has been interpreted to lodge virtually all control of student data in individual schools and institutions of higher education. The department's regulations state that FERPA covers local education agencies, including public schools and higher education institutions, but not state education agencies.[1]

Echoing Campbell, Winnick said that FERPA prohibits these agencies and institutions from disclosing personally identifiable education records without written parental consent, unless the disclosure is covered by a list of exceptions. He observed that "education records" have been very broadly defined to include records, files, and other materials directly related to a student and maintained by an education agency. For example, a case that challenged students' peer review and editing of other students' papers went all the way to the Supreme Court, which ruled that these papers were not education records under FERPA.

FERPA directs the Department of Education to seek voluntary compliance before imposing the sanction of cutting off federal funds—a sanction

---

[1] Currently, FERPA applies to "an educational agency or institution to which funds have been made available under any program administered by the Secretary if (1) The educational institution provides educational services or instruction, or both, to students; Or (2) The educational agency is authorized to direct and control public elementary or secondary, or postsecondary educational institutions" (U.S. Code, Chapter 34, Section 99.1).

to be applied only if the school or school district establishes a policy or practice of making unauthorized disclosures of education records. A 2002 Supreme Court decision clarified that parents and others may not sue a school or local education agency for alleged violations of FERPA. Winnick said that, although the Department of Education has never imposed this sanction,[2] the law has had a "significant chilling effect" on the development of robust state education data systems. He said it is clearly permissible to disclose deidentified data, not traceable to individual students, from state longitudinal data systems. In addition, FERPA does not prohibit creating a state data warehouse, obtaining personally identifiable information from student education records, or using these data to evaluate schools, districts, postsecondary institutions, and programs, including making accountability determinations. State education agency employees and contractors engaged by the state may review and analyze personally identifiable information for these purposes.

Winnick agreed with Campbell that parental consent is not required if a school or local education agency discloses personally identifiable education information to organizations that will conduct research studies for or on their behalf to improve instruction. However, he said, it is unclear what "for, or on behalf of," means and whether a state education data system may authorize such studies. The proposed new regulations do not appear to authorize state education agencies to redisclose data for studies, and, if such authorization were provided, would continue to make such redisclosure subject to the current requirement that a school or local education agency record each disclosure or redisclosure. Winnick asked whether this would mean that the state data system must obtain permission from the school or local education agency, either at the time of redisclosure or at the time when the state first obtains the data (disclosure from the local education agency to the state). He suggested that the proposed regulations be revised to permit recordation of the disclosure by state education officials at the time they make a redisclosure.

Campbell responded that the reason for recording a disclosure is so that a parent or student would know who is seeing their records, and that this was an important issue when FERPA was first implemented. She said that, even if the state were authorized to record the disclosure, a copy of that record would have to be in the school, where the student or parent could see it. Winnick agreed but said that parental requests to see these records of disclosures were infrequent. He suggested that, if there is a parental request, that request be forwarded by the school to the state,

---

[2] Campbell agreed that the sanction has never been imposed, explaining that the local education agencies have responded to threats of funding cutoffs by coming into compliance with FERPA.

so that, when the state authorizes a redisclosure, the state will send that information back to the school. He also said it might not be very difficult, because most the information could be exchanged electronically. Campbell said that it was unclear how frequently parents were requesting records of disclosures, because the department heard of such requests only when the school or school district denied access and parents complained to the department. Winnick did not disagree but said the department's own cost-benefit analysis of the proposed regulations includes low estimates of the number of parental requests for records of disclosures.

Winnick said the proposed regulations would provide a needed change in the Department of Education's interpretation of the FERPA authorization for release of data for studies "for, or on behalf of," education agencies and institutions. Although the current interpretation excludes studies initiated by a research organization, the proposed new rule would authorize release of data if there is an agreement between the education agency or institution and the organization performing the study. However, he indicated that the proposed regulation does not provide for state education authorities to enter these research agreements, thus undermining a key purpose for the state education data systems. Winnick pointed out that the Department of Education's own National Center for Education Statistics, which is subject to the same FERPA provisions as state data systems, has a long-standing practice of entering licensing agreements with third-party research organizations to use student data for research studies.

Winnick welcomed provisions in the proposed regulations that would increase the ability of different state agencies to share education records, which he said would aid creation of data systems that link student data across levels of education, from prekindergarten through postsecondary education. These linked data systems, he said, would be useful for continuous improvement and alignment, to track individual students, and for evaluation and accountability. The proposed regulations would—if they increased the states' authority to record disclosures and enter into studies—provide adequate flexibility in this area, according to Winnick.

In closing, Winnick argued that his proposals would not raise fundamental privacy concerns, because three things would not change: (1) who may receive a disclosure of personally identifiable information, (2) the purposes for which the information may be received, or (3) the fact that the state is maintaining this information. The only real issue, he said, is whether the state—as well as a local agency or individual school—may control the decision and the process to authorize research studies and to disclose data for such studies (and record the disclosure). He concluded that FERPA needs to be reinterpreted, or possibly amended, to harmonize state and federal education policies, rather than to thwart core purposes of state education data systems. Without further changes, he argued, FERPA would frustrate use of state-level, personally identifiable student data for research conducted by organizations other than the state education agency.

In the discussion that followed Campbell's presentation, Marilyn Seastrom (National Center for Education Statistics) explained that one impetus for the proposed new regulations was to address the states' uncertainty about the meaning of "for, or on behalf of." This language, she said, put the state in the position of endorsing that a proposed research study would be useful to the school or local education agency, causing them to be very cautious about disclosing education records for studies. The proposed regulations, she said, clarify that these words simply mean that the study is something the state recognizes as having potential value.

Winnick expressed concern that, although the proposed regulations do allow the state to redisclose student data for a number of purposes, they do not authorize the state to do so for a study. Rather, in implementing the statutory requirement that the study must be "for, or on behalf of an educational agency or institution," the proposed regulations only permit a school, local school district, or educational institution to enter an agreement for a study and release data pursuant to the agreement. He said that it would be legal and quite simple for the Department of Education to clarify that the definition of an "educational agency or institution" includes a state education agency or that the state education agency may enter a study agreement on behalf of schools or school districts in the state.

Responding to a question, Campbell indicated that her office does not formally certify collaborations between education agencies and research groups as compliant with FERPA, but it does review collaboration proposals and provide informal opinions on compliance. Myron Gutmann suggested that, because state longitudinal data systems contain the history of individual students' school participation from kindergarten through higher education, these records could never be completely deidentified. He observed that, if a girl attended a particular school in second grade and another school in a different county in fifth grade, the individual would be obvious as a unique case in the data. Gutmann asked whether any transfer of data from such longitudinal databases to a researcher would constitute a disclosure of personally identifiable information. Seastrom agreed that, because it would be easy to identify one person's unique case, such a transfer would always pose a problem unless the education agency did not allow any directory information at all in the deidentified files. Gutmann responded that he would talk later in the workshop about similar situations, in which researchers are allowed access to data only with strong protections, such as the use of synthetic data or strong contracts between researchers and data holders, such as those used by the National Center for Education Statistics and by his center.

INITIATIVE ON REPORTING STATE EDUCATIONAL PERFORMANCE DATA

Ross Santy (U.S. Department of Education) described several public and private initiatives to use and publicly report education data to improve students' academic achievement. Because these initiatives focus on reporting of aggregated data, representing an entire school, school district, or state, the potential for disclosure of individual student information is small. Nevertheless, as discussed below, the states must make alterations before reporting some subsets of these data files, in order to ensure compliance with the confidentiality requirements of FERPA.

The Department of Education's ED*Facts* initiative aims to access and use student performance data created by the testing requirements of the No Child Left Behind Act of 2001 in order to create a usable national data set (U.S. Department of Education, 2008b). The No Child Left Behind Act requires the states not only to collect and analyze student performance data, but also to publicly report these data in the form of state, district, and school-level "report cards." The report cards must provide student achievement results, both overall and within different groups, such as those with limited English proficiency, those with diagnosed learning disabilities, and members of different racial or ethnic groups. The report cards must also include information on the percentage of students not tested, two-year trend data by subject and grade, each school's status in attaining adequate yearly progress, the professional qualifications of teachers, and other data.

Santy explained that, as the states have increased their public reporting of aggregated educational achievement and school accountability data over the past few years, they have received more requests from outside education organizations and researchers for access to their data in a form that is more usable than the report cards. Many states have responded, posting an annual report card data file on their websites, along with the annual report card. At the same time, the states have received more requests for data from the ED*Facts* initiative. A key component of this initiative was the creation of the Education Data Exchange Network (EDEN) in 2004, as a pipeline to bring state data into the department.

To reduce the burden on the states of complying with these multiple requests for data, the Data Quality Campaign, a national organization dedicated to assisting states in developing and using longitudinal data systems, has collaborated with the ED*Facts* office and independent education organizations to develop the "Coordinated Data Ask" (CDA). In response, state education agencies are beginning to make CDA files available for download on their websites. These files represent a common set of indicators that are often requested by policy analysts and researchers. Reflecting the growing interest in access to these data, the Council of Chief State School Officers and Standard & Poor's (2008) are developing the State Education Data Center, to be an online repository for consistently formatted files of aggregated school-, district-, and state-level education data. The website, created in late 2007, has had a rapidly growing number of hits.[3]

Returning to his description of the ED*Facts* initiative, Santy said that the ED*Facts* reporting tools were introduced in spring 2006. For the first three years of the initiative, states voluntarily provided data to ED*Facts* and, in 2007, the department finalized a regulation requiring electronic submission of data and granting states a two-year transition window. The formal data request to the states was designed to encompass the range of types of data the states were already required to collect, including basic demographics, student performance data, measures of adequate yearly progress, and other data. The states have provided an increasing percentage of the data types and amounts requested by the department over the past three years.

Santy explained that analysts in the ED*Facts* office are currently compiling the data from state education agencies, studying its quality, and determining what is appropriate to be shared with the public. Although they definitely plan to make the national data set public, and they would like to share it with the State Education Data Center in order to reduce the burden of reporting requirements on the states, they have not done so yet. Santy explained that the ED*Facts* office has not made the data public because it has not received a consistent answer from the Department of Education about what constitutes "appropriate" privacy protections. He explained that different offices in the Department of Education have quite different policies and procedures, including licensing agreements used by the National Center for Education Statistics and different data-masking procedures used by the Office for Civil Rights when it makes data sets public and by the Office of Special Education Programs in its annual reports to Congress on implementation of the Individuals with Disabilities Education Act.

Santy outlined several questions about confidentiality facing the states as they cooperate with the Department of Education, the Data Quality Campaign, and the State Education Data Center to share and report their performance data. The No Child Left Behind Act includes provisions designed to protect the confidentiality of student records, directing

---

[3] See http://www.schooldatadirect.org/.

each state to establish a "reporting N size." Under the law, if a cell were to fall below this N size, the state must ensure that the information in that cell would not be publicly reported in a state education report card, in a data file, or in any other way. The data must be redacted, masked, or otherwise controlled. One frequent question from the states is how to respond if one subgroup falls below the reporting N size for a given grade or performance level. Should the state suppress the one subgroup and the total or, alternatively, report only the total data and suppress all of the subgroup data? A more fundamental question is, as growth models are more frequently used to meet No Child Left Behind Act testing requirements, what data should be reported and how will privacy be protected over time?

Santy said that, although the Department of Education currently has no official answer to such questions, officials are now discussing the different approaches used in the department. These discussions address the different uses of the data sets, why they are being made public, and whether different purposes may require different procedures for maintaining confidentiality. Santy described the current situation as "unfortunate," with many different interpretations of how to protect individual privacy of student records, but no common approach.

DATA LICENSING SYSTEM OF THE NATIONAL CENTER FOR EDUCATION STATISTICS

The National Center for Education Statistics at the Department of Education balances confidentiality with research access through a data licensing system, described by Marilyn Seastrom. Seastrom opened by noting that the center uses somewhat different disclosure protections for its own sample survey data than for administrative data owned and maintained by the states. She said that applying the same protections used for sample data to state administrative data sets would reveal the center's approach to protecting the sample data, which includes adding noise and substituting data. Such changes would be readily apparent to state analysts who are familiar with their data and would be inappropriate, because the state data do not belong to the center.

Seastrom explained that the National Center for Education Statistics collects and compiles statistics on education in the United States at all levels, from preschool through adult education. In the process, the center often obtains confidential data about specific institutions and individuals. As required by the Education Sciences Reform Act of 2002 (P.L. 107-279), the center has established confidentiality standards that define individually identifiable information to include "any record, response form, completed survey, or aggregation thereof from which information about particular individuals . . . may be revealed." Such information includes not only direct identifiers, such as a name or social security number, but also indirect identifiers (e.g., place of birth, race/ethnicity, a specific geographic location) that in combination are linkable to a specific individual. She said that a microdata set with thousands of variables for tens of thousands of people could potentially include many unique strings of individual cases across all those items, and this is what the center tries to guard against disclosing.

Seastrom said that the center's mandate to protect privacy and confidentiality is governed not only by FERPA but also by specific confidentiality provisions in the Education Sciences Reform Act and by the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). Section 183 of the Education Sciences Reform Act, known as the Institute of Education Sciences (IES) confidentiality law, states that no person may

- use any individually identifiable information collected under an Education Sciences Reform Act nondisclosure pledge for any nonstatistical purpose, except in the case of terrorism;
- make any publication whereby the data for a particular person can be identified; or
- permit anyone other than the individuals authorized by the director to examine the individual reports.

In addition, both this law and CIPSEA regulations exempt the center's individually identifiable data from the legal process, including requests from the public under the Freedom of Information Act.

Seastrom highlighted two important ways in which the IES confidentiality law differs from more general FERPA requirements. First, it states:

> employees including temporary employees or other persons who have sworn to observe the limitations imposed by this law, who knowingly communicate any individually identifiable information will be subject to fines up to $250,000 or up to five years in prison or both (Class E felony).

Second, the law allows the commissioner of the National Center for Education Statistics to use temporary staff, including federal, state, or local agency employees and private employees, "but only if such temporary staff are sworn to observe the IES confidentiality law." This specific clause, Seastrom said, is the basis of the center's licensing process, providing the legal authority to require people outside the center to "take our oath and have access to the data." Winnick expressed disagreement that these provisions gave the center broader legal authority than the states have to disclose student data to third parties for research studies.

The center's confidentiality procedures guard two types of released data: (1) restricted-use data (the type provided to researchers under license) and (2) public-use data. For restricted-use data sets, the center removes all direct identifiers and either makes confidentiality edits (data perturbation) or restricts cell sizes. Data perturbation techniques directly alter individual respondents' data for some variables, such as blanking out randomly selected records; combining multiple records into a single record; adding random noise; and swapping or switching data. Release of public-use data sets begins when center analysts conduct disclosure limitation analysis of the restricted-use data, in order to determine which records require masking prior to public release. This analysis leads to further data perturbation to create public-use data sets. These changes protect center employees against the disclosure penalties described above, as employees can honestly tell a judge that they did not know whether data they released included individually identifiable information. A disclosure review board made up of technical experts approves data perturbations for restricted-use data files and clears release of masked public-use files. To reduce the need for release of both types of data, the center provides a data analysis system. This analysis software system provides online tabulations in a framework that allows external users to analyze individually identifiable data without direct access to individual data records.

In the data licensing system, center data security staff issue licenses for access to restricted-use data, while contracted security investigators conduct inspections to ensure that the confidentiality provisions of the license are met. The license itself is between the Institute of Educational Sciences, the user, and the user's institution or organization. The licensing system began as a trial in 1991 and was formalized in 2002 following passage of CIPSEA, which allows federal statistical agencies to enter into licensing and contracting agreements. In 2007, the center implemented an electronic system to apply for licensing, and the center expected to issue 900 licenses for access to restricted-use data over the course of 2008.

Individual researchers must apply for a license through an organization based in the United States, submitting a formal request with detailed information about the research project, its objective, an explanation of why the public-use files or the center's data analysis system cannot meet the researcher's need, and other information. Once the application is approved, the researcher must submit the signed license documents, along with a formal security plan and notarized affidavits of nondisclosure, for all those listed on the license as authorized users. The license documents include detailed information about the data, the authorized users, security requirements, and penalties for misuse or disclosure, and they must be signed by both the researcher and a senior official authorized to legally bind the institution. The security plan specifies the exact location where the data are held and used, the physical security of the building and room, and required computer security provisions. For example, the restricted-use data must be loaded and run on a standalone computer, any network devices must be disconnected when the restricted-use data are installed and used on the computer, and the data must be purged and overwritten prior to reattaching to any network.

Seastrom explained that contracted security personnel monitor compliance with all aspects of the security plan and security requirements, which limit data use to a secure room or office, require a password (which must be changed every three months) to log into the computer, limit the data to read-only access, and require the licensee to remove the data either at the end of the project or when reattaching the computer to the network. The license gives center data security officials the right to conduct unannounced, unscheduled inspections of the licensee's site. At the same time, however, the licensee is responsible for ensuring that daily operations comply with the license and security plan, maintaining a file of license-related documents, and ensuring that all authorized users read and understand this file. In addition, the licensee must submit any presentation or publication to the center for disclosure review prior to release, notify the center of any changes to the project or its staff, and close the license after destroying the restricted-use data.

Seastrom closed her presentation with several licensing lessons learned for other agencies and organizations wishing to protect data:

- maintain complete and detailed records of all license transactions;
- use security inspections to monitor minor violations;
- maintain regular contact with licensees, using e-mail and automated features of the electronic license system (e.g., by sending annual reminders for personnel and security updates); and
- automate license closeout reminders.

She observed that, although the center's licensing system was developed to conform to the IES confidentiality law, it can work equally well in other situations. For example, the National Science Foundation operates a licensing system that mirrors the center's system, and both the National Institute on Aging and the National Institute for Child Health and Development have used university data labs to distribute confidential data to qualified researchers through agreements that are similar to the center's license and include onsite security inspections.

In response to these lessons learned, Felice Levine asked whether schools, school districts, and states could use a license to provide researchers with access to education records, and Seastrom agreed that licenses could be used to establish a formal agreement for data access. In response to another question, Seastrom said that the center's licensing system is not subject to review by the Department of Education institutional review board. The department's general counsel determined that, given the center's strong legal requirements protecting confidentiality and stringent penalties for violation of these requirements, it was unnecessary for the institutional review board to be involved in decisions about how the center should protect the data. However, a researcher who accesses restricted-use data through the licensing system may be required to obtain approval from the institutional review board at her or his home institution.

Miron Straf asked whether, under the law, a licensed researcher who discovered individually identifiable information in restricted-use data provided by the center was responsible for protecting that information. Seastrom agreed that, in this case, the burden of protection would lie with the researcher. Straf then asked whether the center might play a role in bringing state data, either virtually or physically, into its system of protections with access through licensing, and Seastrom replied that she believed the center would be open to discussing this.

3

The Value of Education Research Using Student and School Records

VALUE OF LONGITUDINAL STUDENT RECORD DATA FOR RESEARCH

Jane Hannaway (Urban Institute) directs the Center for the Analysis of Longitudinal Data in Educational Research (CALDER), funded by the U.S. Department of Education's Institute of Education Sciences.[1] She described four features of state education databases that make them particularly valuable for analysis and research.

First, because they include unique student identifiers, the databases allow researchers to link individual education records over time in order to develop measures of individual learning gains. Researchers can use these measures to eliminate many confounding variables. In the past, investigators sometimes compared student achievement in a classroom at two different points in time, but the classroom might be made up of different students at the later point in time. Perhaps more importantly, these data allow researchers to address the greatest threat to the validity of many educational studies—the fact that students are not randomly assigned to classrooms and schools. Investigators can use the individual measures of academic achievement and other student characteristics in these databases to statistically adjust for the lack of random assignment.

Second, Hannaway said, some state databases include unique teacher identifiers that allow researchers to link teacher records with student records and track patterns over time. This feature of the databases has

---

[1] See http://www.caldercenter.org/.

allowed CALDER research teams to demonstrate the effect of the teacher on student learning and show how widely individual teachers vary in their effectiveness. This feature has also enabled research on the factors that may account for the variation in teacher effectiveness, such as credentials, training, classroom behavior, and experience. Hannaway argued that these studies are important because they help to clarify which factors do indeed promote student achievement, which in turn has implications for cost efficiency. For example, many school districts provide higher salaries to teachers with master's degrees, but recent CALDER studies indicate that, among elementary school teachers, the presence or absence of a master's degree does not affect student learning gains.

Third, the databases consist of census data, including all students and teachers in the state public education system. This feature of the databases allows an investigator to conduct multiple comparisons. For example, in her recent study of Teach for America teachers in North Carolina, Hannaway was able to compare Teach for America teachers with other new teachers, with all teachers, and with fully licensed and credentialed teachers. She described the potential of the databases for multiple comparisons as "very important" for policy purposes and as a valuable complement to random assignment studies.

Fourth, the databases incorporate historical records, a feature that is critical to understanding the effects of a change in education policy. For example, when studying Florida's A Plus accountability policy, Hannaway expected that the policy would have its largest effects on the lowest performing schools. Contrary to expectations, she initially found that the low-performing schools were less likely to change their behavior than the high-performing schools—suggesting that the accountability policy was not working as intended. However, after analyzing additional data from an earlier period, when another policy targeted many of the same low-performing schools, they concluded that the previous policy had already generated behavioral change in the low-performing schools. Without the historical data, Hannaway said, "you could come to a very faulty inference . . . in policy research."

In response to a question, Hannaway said that the quality of district and state data varies. For example, one North Carolina school district employing a large number of Teach for America teachers provided her research team with data, after a long delay. When Hannaway compared these data with a separate list provided by Teach for America, she found an overlap of only 25 percent. In contrast, some states, including North Carolina and New York, have invested in their data systems, are working with multiple researchers, and have accurate, reliable data, she said.

USING LONGITUDINAL STUDENT RECORD DATA IN HIGHER EDUCATION RESEARCH

Tom Bailey (Columbia University) explained that, as director of both the Community College Research Center, which often analyzes longitudinal student record data from state databases, and the National Center for Postsecondary Research, which conducts more time-consuming and expensive random assignment studies, he sees the value of both approaches. He asserted that it was impossible to carry out "meaningful analysis of student experiences in higher education" without longitudinal data from student records, and that the lack of such analyses limits understanding of higher education. He and other researchers would like to be able track students across higher education institutions in order to address such critical questions as whether community colleges are successful in preparing students to transfer to four-year colleges and how well elementary and secondary schools prepare students for higher education. They would also like to track students within higher education institutions.

Bailey observed that the Family Educational Rights and Privacy Act (FERPA) is only one of several barriers that stand in the way of using education records for these types of analysis. One barrier is the fact that, until recently, most people had faith in the quality of U.S. higher education. Because colleges were assumed to be effective, public debates focused on access, rather than on the quality of higher education. Second, because public funding is based on enrollments, community college administrators often define success in terms of current enrollments, rather than thinking about how to improve student success over the course of their college years. Third, increasing student mobility poses a challenge to measuring the performance of individual higher education institutions. Bailey likened policy makers' current focus on accountability in higher education without attention to individual student progress to General Motors examining its performance without gathering data on car sales over time.

National longitudinal databases maintained by the National Center for Education Statistics have yielded valuable knowledge and understanding of student progress in higher education, Bailey said. The Center's National Education Longitudinal Study of 1988 (NELS 88) collects and maintains data on students over time, including their college transcripts, providing the source for many published studies. However, the data come from limited national samples of all students, including a smaller group of about 1,000 to 1,500 students who have ever attended community college. Given the small size of the national community college sample, it is impossible to analyze student progress in a single state, in demographic subgroups, or in a single educational institution. In contrast, the state longitudinal databases, which include records on every student, are much larger, allowing the Community College Research Center to conduct studies of single colleges, subgroups, and other important topics.

Bailey expressed surprise at how rarely community colleges analyze their own internal student records. Many of the 83 colleges participating in the Achieving the Dream project lack information on such questions as how many and which type of students succeed in developmental education and go on to take regular college courses. Although FERPA poses no barrier to a college in analyzing the progress of its own students, weak information technology (IT) systems pose significant barriers. Community college IT systems are designed to track enrollments once each year and to send these data to the state for reimbursement, rather than to track individual students over time. Although some states and colleges are trying to improve their IT systems, many do not place a high priority on analysis of student progress. According to Bailey, members of community college boards do not understand what kinds of research could be conducted using student records and how that research might improve their educational programs.

Studies using longitudinal data to track student progress over time have already yielded important insights, because they allow researchers to track student responses to educational interventions over long time periods, Bailey observed. For example, a study in Ohio (Purnell and Blank, 2004) found that guidance counseling had strong positive effects on student success in the first two semesters of community college, but these effects had vanished two years later. Studies of developmental education in Florida (Calcagno and Long, 2008) and Ohio (Bettinger and Long, 2008) found that taking these remedial courses had little effect over three years, but greater effects over six years. In another analysis, the Washington State Community College Board found that less affluent students tended to enroll in occupational programs, while their more affluent counterparts more often enrolled in prebaccalaureate transfer programs. The board took these findings to the state legislature and won approval for community colleges to offer applied bachelor's degrees and for expansion of transfer programs.

In another example, Bailey presented an analysis of student progression in mathematics developmental and gatekeeper courses at a single institution. Among students assigned to developmental mathematics courses, 34 percent never enrolled and another 13 percent completed a course but never enrolled in the following gatekeeper course. This kind of information can be surprising to community college leaders, who often focus on improving instruction in individual courses without considering how to ensure that students actually attend classes they are assigned to.

In the few states whose longitudinal databases link elementary and secondary school records with higher education records, it is possible to assess the effectiveness of K-12 education in preparing young people for higher education. A 2006 survey (Ewell and Boeke, 2006) found that only 11 state databases included these links, slowing research on this important topic. For example, dual enrollment programs, in which high school students may take college courses, are growing rapidly, but little research is available on their effectiveness. The Community College Research Center has done a preliminary study of dual enrollment in Florida, because it has one of the few databases with linked records for K-12 and higher education.

Bailey argued that it is also important to link student data among colleges, in order to assess community colleges' effectiveness in preparing students for transfer to four-year colleges. More broadly, the increased mobility of all college students makes linked data across colleges crucial for any analysis of higher education outcomes. Without such linked data sets, education officials must rely on weak measures, such as the Graduation Rate Survey. This measure, which includes only full-time students, is a weak indicator of community college outcomes, since 66 percent of community college students enroll part-time. In addition, the survey excludes transfers and students who register in the spring, and it tracks students for only three years.

To measure community college outcomes more accurately, Bailey and colleagues analyzed longitudinal data on individual college students from the National Center for Education Statistics' Beginning Postsecondary Students study. They found that, although only 22.9 percent of students graduated from the institution at which they initially enrolled within three years, nearly 46 percent graduated from either their original institution or another institution after six years. State education databases allow development of much more flexible accountability measures, in Bailey's view. For example, his center has analyzed three- and six-year graduation rates for different groups of students, including transfer students and part-time students, at Florida's 28 community colleges. While improved accountability measures are important, Bailey said, the real value of the state databases is in allowing more comprehensive and sophisticated analysis of student progress.

In conclusion, Bailey reiterated that research on higher education faces many barriers besides FERPA. Individual colleges and state higher education systems could potentially conduct a great deal of valuable research, but this will require a change of priorities, improvement in their IT systems, and increases in their analytic capabilities. While acknowledging that his center wants to build partnerships with the states, bringing its skills and research priorities, he said it was critically important to increase the states' own skills and priorities.

Responding to a question, Bailey said that Florida and Washington are among the very few states that have linked employment data with education data in their longitudinal databases. For example, researchers analyzing linked data in Washington found that, among community college students who took adult basic education courses, those who continued in other courses and completed least 30 credit hours earned more money later than those who completed fewer credit hours (Prince and Jenkins, 2005). On the basis of this research, the state created a new program integrating adult basic education and occupational skills training.

BENEFITS OF RESEARCH ACCESS TO LONGITUDINAL STUDENT RECORD DATA

Susanna Loeb (Stanford University) opened by discussing how FERPA affects university-based researchers' access to data from individual student records. She said that FERPA allows an education agency to share personally identifiable information from student records without written consent if the disclosure is to "organizations conducting studies for, or on behalf of, educational agencies or institutions for the purpose of . . . improving instruction." The study must be conducted "in a manner that does not permit personal identification of parents and students by individuals other than representatives of the organization" (U.S. Code, Title 34, Part 99, Section 31.6). While this provision has allowed valuable research in education policy and practice, Loeb said, it has been interpreted in very different ways. Some schools and education agencies have shared data, while others have not. In her view, the difficulty of interpreting the law has required both researchers and school personnel to expend substantial effort on compliance.

Addressing the question of why an education agency might want to give researchers access to its data, Loeb said that education policy makers often seek research evidence to inform their decisions. However, most school districts and state departments of education have quite limited capacity to conduct research. Researchers at universities and think tanks can provide the time and some of the expertise needed to make the best use of the information that education agencies have. In addition, outside researchers often have the flexibility to look at medium-run and long-run questions that do not help as directly with day-to-day decisions but can inform better decisions in the future.

The first benefit of allowing access is that researchers have time to compile and analyze data, Loeb said. Because linking and cleaning data from multiple sources is time-consuming, very few states and school districts have done so. For example, she belongs to a team of researchers studying the teacher workforces of New York City and New York state, who have obtained, compiled, checked, and cleaned data from over 10 different sources. From New York public schools, the team obtained data on student demographics and test scores and teachers' years on the job. From New York state, they obtained data on individual teachers, including whether they were certified, their scores on the certification exam, and which teacher education program they had completed. In addition, the team identified the institutions from which individual teachers earned their undergraduate degrees and combined this information with the Barron's ranking of college selectivity to construct variables measuring the selectivity of the college from which each teacher graduated. While the research team had time to devote to this process, it is unlikely that any single education agency in New York would be able to compile all of these data sets. FERPA protections apply both to the individual student data and also to the individual teachers when they were students.

Dedicating this time to accessing and compiling data sets, Loeb said, has allowed her team to conduct several important studies, including an analysis of the impact of the No Child Left Behind Act provision requiring school districts to employ only "highly qualified teachers." In response to the law, the New York City Department of Education eliminated emergency certified teachers between 2002 and 2004, replacing them with teachers prepared by alternative certification programs, including Teach for America and New York City's Teaching Fellows Program (see Figure 3-1). As a result of this change, the average math SAT scores of teachers in the poorest schools increased dramatically. Today, the poorest schools employ higher scoring new teachers than the richer schools.

The second benefit of allowing access is that researchers provide expertise. Although school district and state personnel can often answer day-to-day questions by providing accurate, timely descriptive statistics, outside researchers are able to analyze longitudinal data in much more sophisticated ways. They conduct value-added analyses to assess how much various factors contribute to student learning over time and difference-in-difference analyses to compare patterns in two different time periods. Outside researchers also use a variety of techniques for simulating experiments. In addition, they are using longitudinal data and "putting experiments on top of them," Loeb said. After randomly assigning students or schools (or both) to treatment and control groups, Loeb said, the researchers are not required to gather survey data from the two groups, relying instead on the data that are collected on an ongoing basis in a state or school district database.

Loeb offered two examples of important findings resulting from outside researchers' expertise. First, her team used value-added modeling of longitudinal data to estimate the effect of the "highly qualified teacher" requirement on student achievement (Boyd et al., 2008). They found that

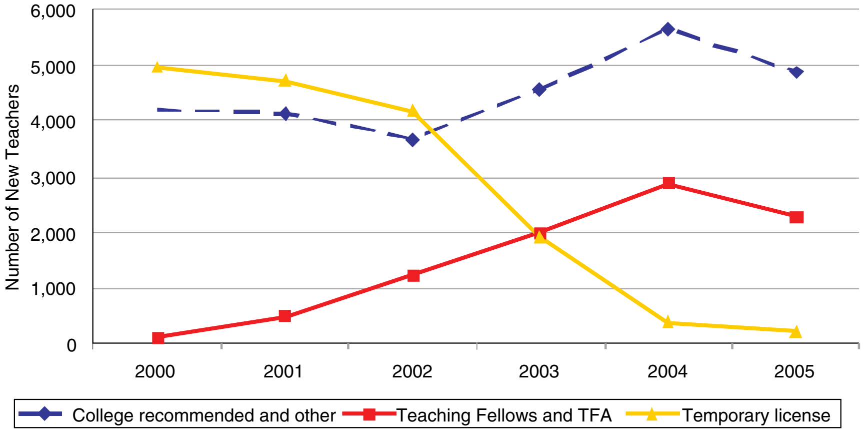

FIGURE 3-1 Number of new teachers in New York City by pathway, 2000-2005.
NOTE: TFA = Teach for America.
SOURCE: Boyd et al. (2008: Figure 6).

the largest increases in teacher effectiveness were in low-income schools, where the weakest teachers were eliminated, whereas the policy had little impact on teacher effectiveness in the richer schools. These improvements in teacher qualifications in the poorest of schools reduced the gap between rich and poor schools in student achievement by 25 percent (see Figure 3-2).

Second, she described studies by Jacob and Lefgren (2004, 2007) of a policy introduced in Chicago public schools requiring students scoring below a specific cut score on a reading and mathematics test to be retained in grade. The researchers used regression discontinuity analysis to compare quite similar students whose scores were below and above the cutoff score—an improved approach over previous studies, which often simply compared the academic achievement of students who were retained with the achievement of other students who were not. In contrast to previous studies, which generally have found that retention has a negative effect on student achievement, Jacob and Lefgren (2007) found increases in measured academic achievement one year later among students who were retained in third grade. However, in comparison to students not held back, these gains vanished by the time the students reached sixth grade.

The third benefit of sharing student record data with outside researchers is that researchers' broad perspective allows them to address questions relevant to long-run policy. For example, Boyd et al. (2005) combined data from college applications to the State University of New York with information from the College Board to describe how close to home teachers

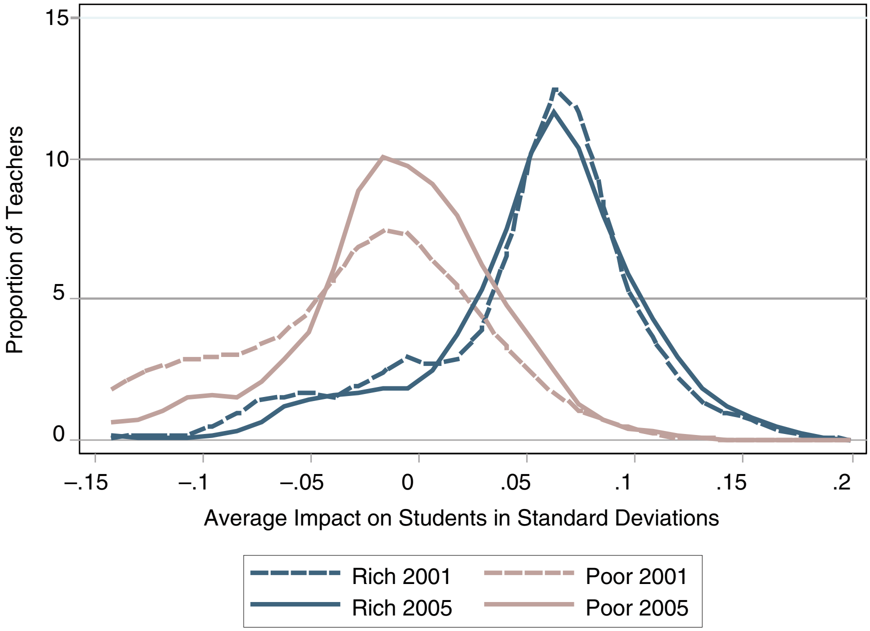

FIGURE 3-2 Effect of all observed teacher qualifications on students in grades 4 and 5 mathematics achievement, most affluent and poorest deciles of schools, 2001 and 2005.
SOURCE: Boyd et al. (2008: Figure 8).

tend to teach. They found that most public school teachers in New York take their first public school teaching job very close to their hometowns or to where they attended college. Teacher candidates coming from suburban or rural hometowns strongly prefer to remain in those areas, rather than teach in urban districts. Their findings have particular implications for the long-term policies of urban districts, which are net importers of teachers. The study suggests that urban districts must offer salaries, working conditions, or student populations that are more attractive than those of the surrounding suburban districts to attract sufficiently qualified candidates.

The broad perspective that outside researchers bring to education questions is apparent in studies that yield important policy information for more than one education agency. For example, an analysis conducted as part of the study of New York City teachers (Boyd et al., 2005) identified differences in the effectiveness of various teacher education programs, as measured by student achievement. The study also identified features of teacher preparation programs associated with greater gains in student achievement—such as providing preparation for teaching practice and developing knowledge of content areas. These findings have implications for other organizations—specifically, colleges providing teacher education—as well as for the New York City Department of Education.

Despite these important benefits to education agencies that share data, Loeb said, researchers often find it difficult and costly to gain access to education data sets. She asserted that the many local school districts, states, and higher education institutions that are interpreting FERPA lack clarity about how to comply. As a result, people in each organization have to think about compliance before providing access. For example, Loeb said, the research team studying the New York state and city teacher workforces had to obtain approvals for the research from over 20 different state and local education agencies and higher education institutions. Obtaining approval from the institutional review boards at 18 different colleges and universities engaged in teacher preparation was nearly a full-time job for one member of the research team, and compiling the data took all of another researcher's time. The process led to a different data-sharing agreement with each organization. The team has a contract to act as an agent of New York State and has signed memoranda of understanding with many school districts across the country, each of which is slightly different from the others. A few districts do not want a formal memorandum of understanding but require the research team to fill out a form.

This process has "huge time costs," Loeb said, partly because schools and agencies are nervous about complying with FERPA. For example, although her workshop paper (Loeb, 2008) includes sample language from a memorandum of understanding with one school district, most of the school districts were unwilling to publicly share their memoranda, because of uncertainty about compliance. Another result of the process is that researchers must work with incomplete and unrepresentative data, because agencies that do not want to share their data use FERPA as an excuse not to provide them. Even agencies that are willing to share data sometimes do and sometimes do not, depending on how much time they have and whether they know and trust the researchers.

Ultimately, Loeb said, the extent of data sharing depends on researchers' ability to develop trust with individual education officials and analysts, which has benefits. For example, her research on the New York teacher workforce was strengthened by extensive discussions with city and state officials. She explained that she is able to access other data from other school districts because she is part of a group at Stanford that gives executive training to superintendents from around the country about the benefits of sharing data with researchers. The bottom line, Loeb concluded, is that there would be benefits to making FERPA "a little bit more understandable" to school districts and state departments of education.

She said that the National Center for Education Statistics' data licensing system was a good model for protecting confidentiality while also providing access. Although the protections are much stricter than those included in her team's contracts and memoranda of understanding with other organizations, it is "much, much easier and less time-consuming for us" to comply, because there is a manual to follow.

VALUE OF RESEARCH PARTNERSHIPS

Barbara Schneider complimented the speakers, observing that CALDER was doing "the most important work on the state longitudinal databases that we have" and that no one was conducting the kinds of analyses of community college education that Bailey's group was undertaking. However, she expressed deep concern about the value and importance of education research. Observing that none of the states has the time or analytic capacity to carry out the types of studies described by the panelists, she said that the real barrier to increased access to state data has been that the researcher has gone in, taken the state's data, and then the state officials never hear from the researcher again—leading to negative feelings about researchers. Schneider called on researchers to establish a new form of relationship with state education officials, including the ideas not only of researchers, but also of the state, and emphasizing the shared interests of both parties.

Praising Loeb's "spectacular" research, Schneider said its results are important, particularly the finding that, at the high school level, the alternative certified teachers are more effective than traditionally prepared teachers, as measured by student achievement. The real question, she said, is what will happen when Loeb and colleagues publish these findings, which reflect negatively on the traditional teacher education institutions that provided data to Loeb's research team. She asked how researchers can go back to agencies and institutions with which they have signed memoranda of understanding to discuss findings that may be negative and, if so, whether the agencies or institutions might pressure the researchers not to publish such findings. She asked about the long-term implications, particularly in light of her call for a new relationship between researchers and education agencies.

Loeb responded that she has observed a change in the way her team interacts with New York City school officials. Five years ago, she said, no one wanted to know about any weaknesses in the teacher workforce, but today all school officials want information on their teachers' "fixed effect" on student achievement, are happy to share that information with the city, and want to know how they can do better. She said that, as university-based researchers, her team retains control over the research information and will publish it. At the same time, however, education officials are likely to recognize errors in the team's data or in its interpretation of the data. Recognizing the value of this expertise, Loeb's team shares draft papers with agency officials, allowing them 30 days to comment, a process she described as "good for them and for us."

Bailey said Schneider had raised a potentially serious issue, as it can create problems if researchers find that an existing policy is ineffective. His team, too, shares draft papers with state officials and discusses the drafts before publication. He said he believes that, in some cases, continued access to data has been limited because of studies reaching negative conclusions and described this as an ongoing problem. However, he emphasized that good access to state education data often is the result of long-term investments by a few states. For example, Florida has held a three-day conference annually for the past 20 years, including all those responsible for sending data into the comprehensive state database, to discuss technical issues. Similarly, the state of Washington has a very good database on community colleges that has been developed with strong political support over 15 years. These states not only have better quality data but also have easier relationships with researchers when discussing such issues as negative findings about education policies. Bailey suggested supporting sustained state efforts like these.

Hannaway noted that, when accepting a grant or contract from an education agency, CALDER always retains the right to publish research results but is flexible about when to publish. The center tries not to blindside education agencies that have provided longitudinal data. In addition, the researchers try not to prejudge educational programs that are still at an "incubator stage." The researchers take time to develop trust with education agencies and to ensure that the researchers fully understand the policy or program they are investigating. Helen Ladd (Duke University) said that, although researchers involved in establishing the North Carolina Education Research Data Center developed trusting relationships with the state and school districts, the center now makes the data sets available to outside researchers, both inside and outside North Carolina. This could have drawbacks if an outside researcher conducted a weak study that would put the North Carolina Department of Education on the defensive. Felice Levine observed that these concerns about publishing negative results, while very important, represented a dimension of conducting responsible, ethical research that is not specific to FERPA.

Focusing more specifically on FERPA, Levine observed that access to personally identifiable student record data is often provided without requiring written consent under the law's exception for studies conducted "for, or on behalf of," an education agency. She asked whether contracts between researchers and agencies reflecting this provision of the law always protect researchers' autonomy to publish their findings. Loeb responded that her team's memoranda of understanding do guarantee the right to publish, as required by the team members' universities. Bailey expressed the view that researchers' right to publish can be guaranteed in the language of the contracts they negotiate, but the real question was whether researchers would be allowed further data access after publishing negative findings about a state or district. His center's researchers have sometimes encountered problems when an individual staff member with whom they have developed a relationship leaves the agency, which has sometimes led to limits on access or lengthy delays before approval of the next data request.

A member of the audience suggested that the take-home message of the panel, including the examples of successful research, was that FERPA and the Common Rule were "really not much of a problem" for researchers. Hannaway disagreed, saying that it was important not to underestimate the costs of obtaining access to these data sets and the "tenuousness" of the relationships researchers had established with states and school districts.

Martin Orland (WestEd) asked whether there were cases in which researchers had tried to gain access to data but FERPA posed a barrier. Hannaway said this had happened in Texas: John Kain, at the University of Texas at Dallas, established relationships and negotiated data-sharing agreements with the state and local school districts, which included confidentiality protections in compliance with FERPA, in the Texas Schools Project. With support from the Spencer Foundation, Kain's team compiled these and other data from multiple sources into a comprehensive longitudinal database with individually linked records on K-12 and higher education and employment outcomes (Kain, 2000). Analyses using the database yielded important findings about student achievement gaps (Kain and Singleton, 1996) and teachers in Texas (e.g., Hanushek, Kain, and Rivkin, 2004). However, Hannaway said, a change in the state's interpretation of FERPA led it to block access to more recent data (Hanushek, 2007). Although the researchers made several efforts to obtain renewed access, including an appeal to the state legislature, they were unsuccessful.[2] Describing the state's decision as "arbitrary," Hannaway called for central guidance on how to interpret FERPA.

Marilyn Seastrom said that the value of developing trusting relation-

---

[2] The University of Texas at Dallas has recently established a state-designated Education Research Center in collaboration with the Texas School Project. According to its website, the new center will assemble, clean, and document deidentified K-12 and higher education data for analysis by the center and will also facilitate secure use by outside researchers in compliance with FERPA (http://www.utdallas.edu/erc/about/ [accessed July 2008]).

ships with the states is not unique to outside researchers. National Center for Education Statistics analysts who create the central core data always send the data for each state back to the state where it originated. This is done partly to verify and edit the data, but also so that the state department of education knows in advance what information will be made public. The center has an "elaborate process" of keeping state education officials involved and informed.

Shelly Martinez (Office of Management and Budget) agreed with Seastrom that these issues of data access are not unique to university researchers. She explained that she participates in the Federal Committee on Statistical Methodology, whose members discuss the use and confidentiality of administrative records across all federal agencies.[3] Martinez observed that there are confidentiality laws similar to FERPA in every field, including health care, in which HIPAA protects individual health records. She said that federal statistical agencies need clear guidance on how to interpret these laws across a variety of situations, because federal agencies are often "in just as tenuous a situation as many of you" when they seek access to state or federal administrative data. Based on her monthly discussions with federal analysts studying nutrition, income, and other topics—all of whom face similar challenges—she suggested developing broad, systematic solutions, as well as addressing the more specific data access challenges posed by FERPA.

In conclusion, Ladd observed that it is important to remember that access to data is sometimes limited by technical weaknesses in state IT systems, not only by FERPA.

---

[3] See http://www.fcsm.gov/committees/cdac/.

4

Reconciling the Access, Privacy, and Confidentiality of Education Data

This chapter discusses models that reconcile research access to education records with the confidentiality requirements of the Family Educational Rights and Privacy Act (FERPA). The first two sections of the chapter describe well-established models, in North Carolina and Florida, that allow researchers to access and analyze state longitudinal data. The next section presents an emerging long-term research partnership that permits University of Illinois at Urbana-Champaign researchers to access and analyze longitudinal data from schools and school districts. The final model is a new collaboration in Michigan.

NORTH CAROLINA EDUCATION RESEARCH DATA CENTER

Helen Ladd provided an overview of the North Carolina Education Research Data Center, which she described as “one of the most productive collaborations between a state department of education and researchers.”

The data center was originally created in 2001 through a memorandum of agreement between the North Carolina Department of Public Instruction and a consortium of researchers at the University of North Carolina at Chapel Hill, and Duke University. The data center is housed at Duke University’s Sanford Institute of Public Policy.[1]

Today’s center had its origins in 2000, when researchers at the two

---

[1] See http://www.pubpol.duke.edu/centers/child/ep/nceddatacenter/index.html.

universities were developing a joint proposal for research into the black-white achievement gap. Realizing they would need large amounts of data, they began discussions with the Department of Public Instruction. The group succeeded in obtaining funding from the Spencer Foundation for two initial studies, for the creation of the data center, and for a colloquium series designed to bring researchers using the data together with policy makers. Although the colloquium series helped to forge ties between state education officials and the university partners, Ladd said, the researchers always emphasized to the state leaders that any research conducted through the center would be independent from the state. Following the signing of the memorandum of understanding in 2001, the Spencer Foundation provided additional support for the data center in 2003, and the memorandum of understanding was updated in 2006. Ladd described the center as a four-way partnership, including the state department of public instruction, the two universities, and the Spencer Foundation (Muschkin and Ladd, 2008).

Ladd explained that the center was initially established to assemble data from the Department of Public Instruction for two purposes—to enable the specific studies of the minority achievement gap funded by the Spencer Foundation and to make the data available to the wider research community. Both purposes have been more than achieved. To date, 93 studies have received data through the center, including 21 projects headed by researchers outside North Carolina.

Ladd argued that the data center's greatest accomplishment has been to overcome barriers to research using the state's education data. In addition to barriers related to compliance with FERPA, the Department of Public Instruction stored administrative data in a format that researchers could not use, and lacked resources to link data on teacher characteristics with data on student achievement, or to create longitudinally matched data over time.

The data center has overcome each of these barriers, by encrypting the data to maintain confidentiality, checking the data for consistency and accuracy, writing user-friendly code books, merging data across sets (e.g., students with teachers and longitudinally over time), and meeting with researchers to explain what data are and are not available. The data center is populated almost entirely with data from the Department of Public Instruction and does not include data from other state agencies. These data are at the district, school, teacher, and student levels. Although some student and teacher records are matched over time, and some teacher and student data are linked, most matching is done by researchers.

Returning to the issue of confidentiality, Ladd said that the memorandum of understanding is very clear on this subject, referring to FERPA and also to the state board of education policy manual. In this memorandum, the state interprets broadly the FERPA provision that exempts disclosure of education records from requirements for informed consent if the records are to be used for studies "to improve instruction" (see Chapter 2). The memorandum governs the process through which the state data are made available. Initially, the data with all identifying information are maintained on a secure server at the Department of Public Instruction. No more than three members of the data center staff—who have been trained in the confidentiality requirements of FERPA and have signed confidentiality agreements—see the data in this form. They encrypt the original data, removing all direct and many indirect identifiers and adding new randomly assigned identifiers. When new data become available from the Department of Public Instruction, the unique encrypted data center identifier is used to link the new record to existing data, and new data center identifiers are assigned to unmatched records. As a result of this process, Ladd said, "I would never see any data with identifiers on it. There are firewalls all around that."

The data center makes these deidentified data available only to researchers employed at a higher education institution or other research organization that has an institutional review board—excluding journalists, advocacy groups, and other organizations that lack these procedures. The data are also available to graduate students who provide letters of support from their faculty advisers, in which the advisers assume responsibility for maintaining confidentiality. To access the data, an investigator sends a research proposal for review by the director and the associate director of the data center. If the proposal is approved, the researcher is required to sign an agreement that includes guarantees of confidentiality and also specifies that the research findings will be shared with both the data center and the Department of Public Instruction.

Ladd explained that the data center process has yielded many benefits for North Carolina's Department of Public Instruction, in the form of useful studies. For example, Bifulco and Ladd (2007) found that students make considerably smaller achievement gains in charter schools than they would have in public schools, and that charter schools have increased racial segregation and minority achievement gaps. Another study, by researchers at Duke University's Nicholas School of the Environment and Earth Sciences, showed that blood lead levels in early childhood are related to educational achievement in early elementary school (Miranda et al., 2007). State education officials are also very interested in a study that found that the incidence of problem behavior was significantly higher among sixth graders attending middle schools than among sixth graders attending elementary schools (Cook et al., 2007).

Ladd predicted that further benefits to the state of North Carolina would emerge from a new collaboration between Duke University faculty and researchers at five other universities. These researchers have access to comparable administrative data from other states, including Texas and Florida. Financed by the U.S. Department of Education through the Center for the Analysis of Longitudinal Data in Education Research (CALDER; see Chapter 3), the collaboration facilitates the replication of analyses across states and shared learning among education researchers.

In addition to this useful research, the data center also generates more direct benefits to the Department of Public Instruction, including providing full access to the cleaned and linked data sets generated by the data center. For example, department officials frequently use data center files on student disciplinary infractions, since they are more easily linkable to student demographic and academic information than the data files maintained by the department. In addition, when the department contracts with other state agencies to carry out research, the data center supplies files to these agencies at little or no cost. Over time, the department's internal research capacity and productivity have both risen. The presence of the data center also allows the department to refer requests for data from outside researchers to the center, eliminating the costs of responding.

Ladd emphasized that the research community's data needs differ from those of the Department of Public Instruction. Researchers value longitudinal data over a long time period, in order to estimate models of educational activities and outcomes. In contrast, the Department of Public Instruction sometimes needs the most recent data and data with identifiers. Because of these differences, it is important that the department continue to develop its own data management capacity, Ladd said. Noting that the department had received a grant from the U.S. Department of Education to develop a longitudinal data set, she said it was important that the department move forward on this project internally, even though the data center has been working to create a longitudinal data set. She said that the data center has already held useful conversations with the state about this project, predicting that the evolving relationship between the state and the data center will be productive. Ladd observed that it would take the department two or three years to develop its longitudinal data set and that, unlike the data center's data set, it would not include historical data.

Finally, Ladd noted that the data center and the Department of Public Instruction share the goal of expanding the developing K-12 longitudinal systems to include data on postsecondary education. Through its partnership with the department, which has begun discussions with the state higher education system, the data center hopes to establish the necessary mechanisms for sharing data, ensuring confidentiality, and providing researcher access to information that would promote policy-enhancing research at all levels of education.

In response to a question about the memorandum of understanding, Ladd said that it was helpful during the negotiations to explain to the North Carolina officials that Texas had a similar arrangement to share data with researchers at that time (see Chapter 3). Robert Boruch called for increased sharing of memoranda of understanding to reduce the burden of negotiating each agreement separately and to encourage more states to develop and share data sets.

FLORIDA'S EDUCATION DATABASE

Jeff Sellers (Florida Department of Education) provided an overview of Florida's comprehensive data system. He explained that the state developed the Florida Education and Training Placement Information Program in the 1990s, in order to assess the effectiveness of state educational programs. The system used social security numbers to link individual student records at all levels of education and training to other state data sets on employment and wages, public assistance, incarceration in the state prisons, military enlistments, and other life activities. Analysts used the linked data to answer questions about what happens to students after they graduate from or leave educational programs; for example, one study examined the relationship between high school test scores and receipt of public assistance.

Lessons learned in developing this program were applied to develop a central state data repository in 2000, using more current technology. This central data warehouse is administered by the Florida Department of Education but relies on administrative data from a variety of state and federal agencies. It includes individual assessment results from prekindergarten through community college, including scores on teacher certification assessments, as well as other individual student and staff data from all levels of education, prekindergarten through state university. Since 2004, state analysts have drawn on the warehouse to create several products. Emphasizing that state officials and systems analysts learned as they developed the system, Sellers noted that one of the first products, a series of research extracts, was created in partnership with outside researchers who reviewed the data sets and provided feedback on their accuracy and usefulness. Local researchers from Florida State University also helped the state analysts to study and understand the longitudinal data they had assembled. The next product was a series of data marts, which are made up of data files aggregated by subject, such as enrollments or assessments. They are designed to provide increased access to the data, while protecting confidentiality in compliance with FERPA, Sellers explained.

Aggregating these subsets of the huge volume of records speeds system performance and allows faster responses to specific queries.

Sellers described the state's approach to making the data anonymous. As the data are extracted from the various sources (assessment data, preK-12 student data, etc.), they take two different paths (see Figure 4-1). Data on enrollments, assessments, attendance, and the like follow one path into the warehouse, and student and teacher data follow another path, in which identifiers (social security number, name, birthdate, and other student and teacher identifiers) are stripped off. The data enter a "black box" in which each individual or institution is assigned a unique student or teacher identification code, known as a "data warehouse internal identifier." This anonymous identifier is then relinked with the data that followed the other path, creating a new, anonymous record, which, in turn, is loaded into the data warehouse for storage. Sellers emphasized that this process of matching a new anonymous identifier to each student or teacher is completed only once, mentioning that the group uses a similar technique to make school information anonymous. Data warehouse managers ensure that each unique individual or institution retains the originally assigned identifier throughout the loading and extraction of data.

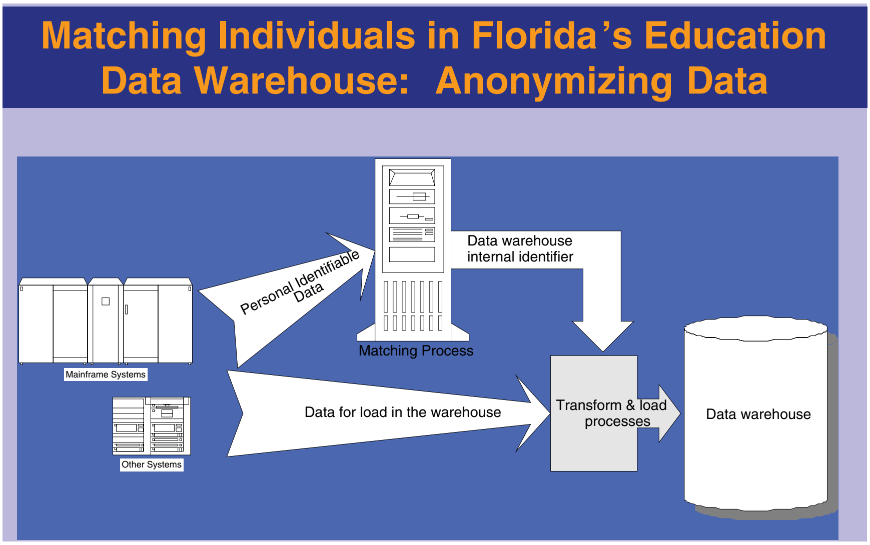

FIGURE 4-1 Florida's approach to making data anonymous.
SOURCE: Sellers (2008).

The integrated longitudinal data in the warehouse are useful for many different purposes. Administrators in the Florida Department of Education use the data to inform decisions about funding, class size planning, and other matters and to meet the federal accountability and reporting requirements. The data marts produced over the past two years, which report and present the data, have been used to provide feedback reports to high schools and community colleges. In response to state legislators' questions, analysts are creating a data mart focusing on the effectiveness of the variety of state-funded teacher preparation and certification programs (including alternative certification programs). Using an approach similar to the one Susanna Loeb used in New York, the state analysts link teacher identifiers with student identifiers and also look at the teacher's educational record when she or he was a university student, to assess the effectiveness of alternative preparation programs based on student performance.

The data are also useful for research, Sellers said. The state of Florida has established research partnerships with the Center for the Analysis of Longitudinal Data in Educational Research (CALDER; see Chapter 3) and has collaborated with the Community College Research Center described by Thomas Bailey to study community college financial aid and high school–community college dual enrollment programs (see Chapter 3).

Sellers concluded his presentation by highlighting key lessons about what has worked in Florida:

- To the extent possible, build on existing systems and expertise. Sellers observed that many states have administrative data, but the key question is how to leverage these data to study and inform policy. For the 27 states that have received federal grants to build longitudinal data systems, he said, the challenge will be linking the data so that students can be followed over time in school and beyond, as they enter postsecondary education or the workplace, receive public assistance, or have other life experiences.
- Pursue opportunities to provide service and share information. Sellers noted that he usually delivers presentations to "people who have the data, not people who want to use it." In his talks, he encourages the data managers to respond to outside researchers' requests by considering how the state may leverage the request. Because his agency lacks the resources needed to evaluate current or proposed future education policies, outside research can be very valuable. At times, state officials will ask the researcher to slightly modify the research plan or add a component related to a specific question in order to gain more from the research. Sellers suggested that this could be a selling point for researchers as they approach

states looking for data. The researchers could propose to help the state evaluate some of its programs and policies in exchange for access to the state's data.

- Exceed all requirements dealing with confidentiality and restricted release. This is important not only because of the legal requirements in FERPA, but also to address how the public, parents, and the media perceive the collection and use of individual data and to counter possible future references to the state government as Big Brother.
- From a development perspective, it is really never over. The answer to any single question about education policy often raises five new questions that need to be addressed.

In discussion, Sellers said that his agency does not have enough staff to respond to all requests for access to the database. He observed that the queue of researchers seeking access was growing longer. In the future, he said, the state would like to create a center similar to the North Carolina Education Research Data Center—a "virtual sandbox." Ideally, the state could give the key to the sandbox to a qualified, approved researcher, moving to a self-service model that would eliminate the need to respond to each individual request.

DEVELOPING LONG-TERM RESEARCH PARTNERSHIPS IN ILLINOIS

Lizanne DeStefano (University of Illinois at Urbana-Champaign) explained that she been involved in balancing researchers' need for access to education data with protections for individual privacy and confidentiality for over two decades, including many years as chair of the university's institutional review board and in her current position as associate dean for educational research. Observing that "we live in interesting times," DeStefano argued that longitudinal studies of individual student performance over time are critical for responding to the accountability requirements of the No Child Left Behind Act. Because of this, she said, local and state education agencies are now more motivated to find solutions that enable research while also protecting student confidentiality.

DeStefano outlined different phases in the relationship between University of Illinois at Urbana-Champaign education researchers and local school districts. In the early 1990s, local schools and districts became increasingly unwilling to respond to many different ad hoc requests for data from university researchers. They viewed researchers as people who came into the schools, took data, and left without providing anything of value to the school or district. In addition, these small schools and districts often lacked the time and staff to respond to individual requests that they create anonymous data sets.

To overcome this problem, the university created the Office of School-University Research Relations, a single point of contact for researchers and schools. School officials can call this office to receive assurance that any research proposed for or under way in their schools has been approved by the university's institutional review board and that the researchers involved have undergone criminal background checks and received training in research ethics and procedures. School officials can also ask questions about the research and its findings and implications for school policy. The new office was successful for many years, facilitating about 150 research projects in schools each semester.

However, in 2004-2005, as school officials grew concerned that spending time with outside researchers was distracting them from improving instruction in order to meet the requirements of the No Child Left Behind Act, they again grew reluctant to sponsor research. At the same time, responding to the federal requirements had increased their awareness of value of data analysis, evaluation, and research. These changes led the university and the school districts to a new phase in their relationships, establishing long-term research partnerships based on common interests and a shared commitment to school improvement. The new partnerships involve many different school districts in the Urbana-Champaign area and are supported by a new Center for Education in Small Urban Communities.[2] DeStefano noted that the university has not yet developed a long-term research partnership with the state of Illinois and only recently signed the first memorandum of understanding governing access to state data.

Although there are many research partnerships, each includes several common elements. First, it is based on a negotiated long-term research agenda, developed through frank discussions among university administrators and faculty and school district representatives. DeStefano observed that, although she had initially feared that faculty interests would differ from the districts' interests, she found quite a bit of overlap. These discussions and negotiations led to a list of four areas in which research is critically needed, and the university provides funding and fellowships to encourage faculty to conduct studies in these four areas. In addition, the university commits to sharing the research findings in a form that the school or school district can easily use and apply.

Second, each partnership deploys similar strategies for informing parents and students about research activities. The partnerships use "robust and effective" procedures for disseminating information and obtaining

---

[2] See http://www.ed.uiuc.edu/smallurban/.

parental consent, when consent is required, DeStefano said. For example, parents can view survey instruments, protocols, and research summaries, and they can also call a toll-free number if they have a concern or question about research under way in their schools.

Third, each partnership is supported by cross-training of school district personnel, families, and researchers on research ethics and compliance with FERPA. This training develops shared understanding and undergirds the fourth element of the partnerships—formal data-sharing agreements specifying that the school districts will maintain and allow regular access to deidentified longitudinal data sets in specific areas. These formal agreements include common interpretations of FERPA and options for compliance. For example, when informed consent is required for disclosure of education records, the relationships the university routinely develops with families make it relatively easy to obtain signed consent forms. More often, the school district is allowed to disclose deidentified data without informed consent under the FERPA exception for "organizations conducting studies for, or on behalf of, educational agencies or institutions for the purpose of . . . improving instruction" (see Chapter 2).

In addition, a memorandum of understanding with each school or district partner allows the university researcher and the district to obtain approval for the research from the University of Illinois at Urbana-Champaign institutional review board, which has been expanded to include representatives of the school districts. No approvals from other institutional review boards are required, saving both the researchers and the school districts time and money.

The final element of each partnership includes measures to strengthen the research capacity of the school district. The university provides resources to strengthen district information offices, including technology, expertise in encryption and security, shared servers, and student interns with background in information technology.

In conclusion, DeStefano said that, although some faculty members had been concerned that the efforts to develop new research partnerships would constrain their research agendas, this has not happened. At first, the negotiated research projects were very focused on students' performance in reading and mathematics, because this is what the schools wanted. However, after three years of experience, including routine meetings with faculty members, school officials now recognize the value of broader research, including investigations of social, emotional, and behavioral questions. For example, one current study is investigating student health and obesity. Responding to a question, DeStefano explained that school representatives agreed to a broader research agenda partly because they had received a good payoff from the original projects focusing on analysis of reading and mathematics performance, including "tables and graphs and charts and reports that they could have never generated." She emphasized that university researchers routinely share reports and studies with school officials.

In response to another question, she said that, with a few exceptions, the university does not provide financial or other incentives for parents to sign informed consent forms. Instead, the university aims to educate and inform parents about research going on in their children's schools, through a newsletter, a series of public forums, an annual conference cohosted by the university and the school districts, and the toll-free number mentioned earlier. As a result of these efforts, when a student brings home an informed consent agreement, the parents are more likely to be aware of the research and are more likely to read and sign the agreement. Over the past five years, the response rate when researchers send out consent agreements has increased.

A NEW COLLABORATION FOR DATA SHARING IN MICHIGAN

Barbara Schneider began by identifying several differences between her new model of collaboration in Michigan and the North Carolina Education Research Data Center. She explained that the Michigan collaboration was developed much more recently, through a subcontract with the Regional Education Laboratory-Midwest, which in turn is funded by the U.S. Department of Education's Institute of Education Sciences. The subcontract has two goals (Schneider, 2008):

1. To demonstrate the feasibility of assisting state education agencies in leveraging existing state administrative record data to provide an empirical basis for developing education policies and practices and
2. To provide technical assistance to the state of Michigan, addressing questions initiated by the state—and to document and publicize the technical assistance process as well as any unique analytic problems in working with the state administrative records.

One continuing challenge, Schneider said, is that in Michigan, the Center for Educational Performance and Information (a unit of the state budget office) collects and maintains education data, rather than the Department of Education. Schneider said that different offices in the center allow her research team (including faculty, postdoctoral fellows, and Michigan State University graduate students) to access and analyze administrative data files in response to questions they jointly identify. These questions and their potential answers are constructed to inform state education decision making and have potential budgetary consequences.

The research team has published three reports, including a background survey and two technical reports to the state. The survey yielded responses from education officials in seven Midwestern states about the types of data and data analysis they would find most useful (McDonald et al., 2007). The researchers learned that state education policy leaders are very interested in developing longitudinal data systems with linked student, teacher, and school data. However, most states lack the personnel and the capacity to develop such systems at present, due to their very limited resources.

Before discussing the team's first technical report, Schneider described the context supporting the analysis. Because the demonstration project was specifically designed to make the collaborative process transparent, a member of the research team took notes at every meeting; all team activities were documented in a historical file; and all data analyses, codes, and procedures were recorded. As required in the subcontract, the team identified unique analytic problems of dealing with universal data, rather than a survey sample. For example, to analyze differences in data on 98,000 teachers, the team used gamma statistics, rather than the t-test commonly used to analyze differences in means between two sample groups.

Schneider outlined three ground rules underlying the researchers' relationship with the two Michigan agencies. First, the questions were to be generated by the agencies. Second, the analyses would be designed and conducted as an iterative process, which led to modifications at each meeting. Third, the research team would document the technical assistance process, to enable an evaluation of the potential replication of this model.

Turning to the issue of confidentiality, Schneider explained that the research team is able to access the Michigan data under the FERPA provision allowing disclosure without prior consent "to organizations conducting studies for or on behalf of educational agencies" and because the research team members are designated as agents of the state (see Chapter 2). In addition, the team members have obtained approval from the institutional review boards at their respective universities and at the state level. In response to a question, Schneider said that confidentiality protections are very important, and that she and other professors teach their graduate students this. For example, she requires her graduate students to practice following the procedures required by the National Center for Education Statistics' licensing process (see Chapter 2). Schneider emphasized the need to inform the research community about confidentiality protections.

The team's first analysis of Michigan data originated in a meeting that included the researchers and staff of the two state agencies, organized by Margaret Ropp, director of the Center for Educational Performance and

Information. The purpose of the meeting was to discuss priority issues that could be addressed through analysis of the existing state data. In preparation for the meeting, Schneider's team created a series of table shells illustrating the types of analyses that could be conducted using the state's data. When they saw these tables, she said, the state analysts "got really excited," because, although they have doctorates and are very familiar with the data, they lacked time to think about questions that the data could answer. She observed that "everybody's eyes lit up" and the meeting continued until 6:15 on a Friday afternoon.

At the meeting, participants discussed Michigan's recently approved "merit curriculum," requiring high school students to complete four years of English and mathematics, three years of science and social studies, and two years of foreign language in order to graduate. The state superintendent of schools wanted to know whether schools across the state had enough qualified teachers to teach the required subjects. To address this question, the state gave the team access to a large file of teacher data. With the help of a postdoctoral student, Schneider was able to "unstack" the file into individual teacher records. When team members verified their individual records related to thousands of teachers against a Standard & Poor's database of Michigan teachers, they found only a small discrepancy. The team then linked these records with other national and state data and analyzed the linked data sets.

Schneider observed that, as the research team and agency personnel began to work together, they built trust. She said that all of the key ideas identified in her book on relational trust (Bryk and Schneider, 2002)—respect, competence, integrity, and working for the common good—were realized over the course of the project. For example, she promised to deliver a report within a month of the first meeting with state officials, in order to counter the view that research "takes forever." To meet this commitment, the five team members worked daily to produce a draft, which they reviewed and discussed with agency officials. After revisions, the team delivered a final report.

Schneider said that her graduate student, who was familiar with high school scheduling, developed a demand formula to answer the superintendent's question about whether schools across the state had enough qualified teachers to deliver the merit curriculum. She said that several statisticians had described this formula as "the most simple, elegant way to figure out how many teachers you need to teach the merit curriculum in your school." The formula is designed to adjust for changes in enrollment size, increases and decreases in class size, and changes in the number of courses teachers are required to cover. Applying the formula to the large teacher data set, the team found that only 14 schools had an undersupply of qualified teachers in all four required subjects, but, when considering each subject and grade level, the undersupply of qualified teachers could potentially affect 72,000 students.

The team's findings on teacher supply and demand were summarized in a second technical report to the state (Keesler et al., 2008). The state officials welcomed the report, using it to target professional development courses and funding toward schools with an inadequate supply of qualified teachers. They have also raised new questions that the research team is currently addressing. At the same time, Schneider and Margaret Ropp, director of the state Center for Education Performance and Education, are sharing the report's findings and methods, including the formula, throughout the Midwest region (Ropp et al., 2008).

In conclusion, Schneider said that the benefits of the project include a trusting, open relationship with state personnel, encouraging collaboration across universities, openness and sharing of information, and working "with a fabulous group of professionals." The greatest challenge, however, is that the state, on the basis of its interpretation of FERPA, has not provided access to linked longitudinal student and teacher data, as the research team originally requested in 2006. Schneider described FERPA as "the shield that stops us and the barrier from getting to the places where we want to be."

Finally, Schneider said that the project is at a much earlier stage than the databases allowing research access in Florida and North Carolina. The team is currently in the process of creating a research collaborative that will allow researchers across the state to access the teacher file and other files, and this has raised questions about who should warehouse the data. Currently, she said, the state of Michigan maintains the data, although both Michigan State University and the University of Michigan have proposed to warehouse the developing data sets. The team is also working on another technical report and collaborating with the state to study several new issues.

5

Reconciling Access and Confidentiality in Federal Statistical and Health Data

This chapter focuses on reconciling research access to administrative data with privacy, confidentiality, and consent requirements in health care and other sectors outside education, as well as considering the implications for education research. The first section describes the Census Bureau's approach to statistical use of administrative data and outlines options for allowing researchers to access data sets while protecting confidentiality. The second section includes an overview of data access and confidentiality issues and further discussion of options for reconciling these issues. The third section discusses the impact of the Health Information Portability and Accountability Act (HIPAA) on health research using medical records, and the fourth section outlines concepts for a data stewardship entity that could potentially facilitate health research. Finally, the chapter summarizes an extended discussion of the implications of experiences in these other sectors for education research.

INTEGRATING ADMINISTRATIVE DATA INTO CENSUS BUREAU PROGRAMS

Gerald Gates explained that his years at the Census Bureau as a privacy officer and earlier as an administrative records program officer had made him aware that privacy is "the key issue" in obtaining access to administrative records and sharing them with researchers. On the basis of his understanding of laws, regulations, and current practices, he outlined three fundamental principles for statistical use of administrative records. First, individuals must be informed of the uses of their personal information and given the ability to control such uses. Second, he argued, administrative data can be shared for statistical purposes without consent, provided the data are protected from nonstatistical uses, echoing a point Straf had made earlier (see Chapter 1). Third, federal agencies must provide effective data stewardship, ensuring both appropriate protections and optimal use (Gates, 2008).

Moving to more practical issues, Gates observed that it is "hard to get these data," requiring negotiations among lawyers, program managers, policy advisers, and institutional review boards in order to reach agreements on data sharing. He noted that any violation of confidentiality protections negatively affects all parties, including an administrative or a statistical agency that shares data with a researcher. However, the parties are not equally liable for protecting confidentiality. Depending on the arrangement for sharing of data, the researcher may not be liable, but the federal agency is always liable—which may make an agency reluctant to provide access. In addition, news reports about breaches of security in federal data systems (e.g., Lee and Goldfarb, 2006) raise concerns among the public and in Congress and put pressure on agencies to protect, rather than share, their data.

Legal and Policy Support

Both law and policy support the use of administrative data for statistical purposes, Gates said. The law not only authorizes the Census Bureau to acquire administrative records, but also goes further to state that the bureau must use such records "to the maximum extent possible . . . instead of conducting direct inquiries" (U.S. Code, Title 13, Sections 6, 9, and 23). The law protects administrative information that is used for statistical purposes from being reused for administrative purposes. The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002, another important law, requires uniform confidentiality protections among federal agencies that collect information for statistical purposes. Prior to this law, Gates said, agencies protected this information under a variety of laws and regulations—some more ironclad than others.

Many policy studies also support the use of administrative records for statistical purposes. In a key report, the congressionally mandated Privacy Protection Study Commission (1977) defined the concept of "functional separation" between use of individual information for statistical purposes and for administrative purposes. More recent reports (e.g., National Research Council, 1993) support the use of administrative records for statistical purposes in ways that protect individual privacy and data confidentiality.

Uses of Administrative Records in Statistical Programs

Gates said that administrative records—such as the tax records the Census Bureau obtains from the Internal Revenue Service—can be useful to statistical programs in several different ways:

- to assess population coverage in surveys;
- to assess the nature and impact of survey nonresponse;
- to aid survey methodologists in understanding the nature and extent of sampling error;
- to improve survey data editing and imputation;
- to improve questionnaire design;
- to make improvements in survey sampling frames;
- to improve simulation models for policy evaluation and review;
- as a source for economic survey sample frames;
- as measures of migration for producing population estimates between censuses;
- as a source of information about income, poverty, and health insurance at the substate level; and
- to investigate social, economic, demographic, and occupational differentials in mortality.

Recent Census Bureau Data Linkage Activities

Gates described several recent data linkage activities at the Census Bureau. Analysts began developing the Statistical Administrative Records System (StARS) before the 2000 census, collecting information from five agencies that replicates the answers to questions on the short form of the census. Originally developed as a low-cost alternative to improving within-household census coverage, the new records system has improved the bureau's demographic information, which will enable improvements in the demographic data collected in the next census. The new system also provides more up-to-date information, such as very current change-of-address data from the U.S. Postal Service, making it a very valuable resource.

The Census Bureau launched the Longitudinal Employer Household Dynamics Program in 1999 to integrate census, survey, and administrative records data on workers and employers, including state unemployment insurance wage records. This data system provides a detailed, comprehensive picture of workers, employers, and their interaction in the national economy. While offering unprecedented detail on the local dynamics of labor markets, the data program maintains confidentiality through advanced confidentiality protection methods (Abowd et al., 2005).

Another recent Census Bureau program to link data sets is the Medicaid Undercount Project. This data system includes information from the Current Population Survey, the National Health Interview Survey, and the Medicaid program. Its goal is to examine "perplexing" discrepancies between estimates of Medicaid enrollment from population surveys and enrollment counts from the Medicaid program's own administrative data.

Protecting Privacy and Confidentiality of Administrative Data

Gates said that the key challenge in obtaining administrative data for statistical activities and research access is to protect privacy and confidentiality. He distinguished between privacy, which he defined as an individual's right to control the use and disclosure of information about himself or herself (Fanning, 2007), and confidentiality, echoing a point made earlier by Miron Straf (see Chapter 1). Collecting and using social security numbers, which play a key role in integrating administrative data sets, raise privacy concerns. To address such concerns, the Census Bureau and other federal agencies convert the social security numbers to protected identifying keys that cannot be decoded except by a handful of persons who know the code. After data sets are linked, social security numbers and names are removed and replaced with these keys.

Public awareness heightens the challenge of protecting privacy and confidentiality, Gates said. For example, in 1999, the privacy commissioner of Canada effectively shut down a major data linking project on the grounds that it had not been sufficiently publicized. In announcing this decision, the commissioner observed that, although the agency in charge of the project (Human Resource Development Canada) had not tried to hide its effort to collect and merge individual information, Canadian citizens remained unaware of how much information was being gathered about them and the extent to which it was being shared with others (Gates, 2008). Gates said it is important to acknowledge and address people's concerns about merging and reusing various sources of administrative data.

Maintaining the confidentiality of individually identifiable records poses a greater challenge today than it did in the past for two reasons, Gates maintained. First, policy makers and researchers in education, health, and other fields are demanding detailed, individually linked data sets. Second, when such data sets are made public, people have access through the Internet to many more sources of data that could be used to identify individuals in the data set.

Options for Providing Access to Administrative Data

Gates presented several options for providing safe, useful access to administrative data. He explained that it was important to consider the range of options, because no single option will meet the needs of every data producer and every data user.

Options for Public Use Without Restrictions

Traditionally, the Census Bureau and other federal statistical agencies have made survey data and some types of administrative data available in the form of "public-use microdata systems," and these data sets continue to meet the needs of most researchers today, Gates said. To protect these data against an intruder who might try to identify one or more individuals, the public-use microdata systems include only a sample of records. For example, for the decennial census, the largest available national public-use microdata file includes only 6 percent of the population. In addition, bureau analysts remove all direct and indirect identifiers, restrict the amount of geographic information shown, remove outliers, and take other steps to reduce the risk of disclosure. The advantages of these data systems include availability to the public with no limits on use and easy analysis using most software. At the same time, public-use microdata sets have several limitations, due to the restrictions on geographic information, the removal of outliers—which are often the most interesting data—and other confidentiality protections.

Gates explained that, for many administrative data sets, further protections are needed to protect the confidentiality of individual information. One option is to develop synthetic data. Synthetic data sets have several advantages. They are designed specifically to protect administrative data, they can be made accessible to the public or to researchers without restrictions, and they are easy to analyze using most software. However, these data also have limitations. Because synthetic data sets are customized to meet the needs of specific groups of users, they will not satisfy all researchers. In addition, research to date has not yet demonstrated the quality and usefulness of synthetic data for a wide range of different types of applications and analyses.

Options for Restricted Use

Restricting use of administrative data sets provides another layer of protection, Gates explained. The Census Bureau pioneered one option to restrict use—the Research Data Center. After establishing the first center in Boston in the 1990s, the bureau expanded the network and today operates nine such centers. At these centers, users may directly access administrative data sets that include everything but direct identifiers (name, social security number). The data include outliers, and the user may link the data to external data sets. However, these centers also have limitations. The researcher or other user must go first go through a complex approval process and then relocate to the regional research data center, limiting possibilities for collaboration with other researchers.

Another option for restricting access to administrative system data is through a licensing system, such as the system operated by the National Center for Education Statistics (see Chapter 2). This option has the advantage of allowing the licensed researcher to directly access data sets at her or his own institution, using the researcher's own software, and facilitating collaboration with colleagues at the institution. The limitations of this option include the possibility of losing one's license if an onsite inspection by the licensing agency finds violations of the confidentiality protections or other elements of the licensing agreement. In addition, the license agreement typically does not allow the researcher to link the licensed data to external data sets.

Remote Access Options

Gates said that the option of providing researchers with remote access to statistical data is growing rapidly. In this option, the researcher submits programs to an intermediary, which applies the programs to restricted-use data and provides the results and tabulations to the researcher. For example, the National Opinion Research Center has created a "data enclave," with data sets from the National Institute of Standards and Technology and other sources and conducts analyses of these data at the request of approved researchers.[1] The National Center for Health Statistics provides access to approved researchers by e-mail through the Analytic Data Research by E-mail (ANDRE) system and has also created a virtual research data center.[2] Remote access options have the advantage of providing easy access to administrative data sets by e-mail or the Internet. However, the types of analysis possible are limited by the specialized software housed on the servers of the data enclave. In addition, outliers in the data sets have been removed to protect against disclosure of individually identifiable information, and some data enclaves require users to pay a subscription.

A final option for providing access to administrative data is to obtain informed written consent from the individuals whose administrative records are sought. Consent has the advantage of putting the individual

---

[1] See http://www.norc.org/DataEnclave/.

[2] See http://www.cdc.gov/nchs/r&d/rdc.htm.

in control and potentially allows the most flexible access to, and use of, individual record data. The limitations include potential bias in the data set, because some individuals will not give consent for use of their records, and it may not be possible to locate other individuals. When using this option, Gates said, it is critically important to test different formats for informing individuals of the proposed uses for their record data and requesting their consent.

Gates concluded with several key points from his paper (Gates, 2008). First, administrative records serve many important research uses; these uses are supported by law and facilitated by the advanced technology and methods for linking records available today. Second, although all parties who have access to administrative records have a great incentive to protect confidentiality, they are not all equally liable for any possible breach of confidentiality. Third, access is affected by increased public scrutiny of privacy protections, which leads to development of new data stewardship principles and policies. Gates explained that agencies sometimes apply new policies and procedures in response to privacy violations "in order to survive."

Finally, he observed that the variety of new options for access reflects the degree of control the agency holding the records is willing to relinquish to permit specified uses. Gates emphasized that agencies will provide access if they are comfortable that their requirements are going to be met, leading to a "staged approach," rather than trying to provide access to all data for all users. Instead, the agency considers the needs of particular groups of data users, provides only the amounts and types of data needed, and imposes restrictions on use of these data.

Responding to a question, Gates said that language is very important when talking about the complex issues involved in maintaining confidentiality of individual information. He said that the key point to convey, when talking to the public about using their personal records for statistical purposes, is that a federal agency will not use personal records to make administrative decisions, such as to determine one's level of social security benefits, and that the information will be merged with information from other individuals. He said that the Census Bureau had done some research to find that, when people were asked to give consent for use of their information "for statistical purposes," they did not understand the term. People were more comfortable when asked to give consent "for statistics," because they think of statistics as numbers, Gates said. In response to another question, Gates agreed with Robert Boruch that studies of the informed consent process are important, stating that "the statistical community needs to acknowledge that consent is an important issue."

Felice Levine highlighted a key question related to informed consent.

She said that, if researchers ask for informed consent when an individual takes a test or gives blood, they must consider what the consent is for. This question becomes important, she said, when a researcher asks an institutional review board for a waiver of consent. The board will look at the proposed research and consider whether the original consent—for its intended purposes—would be compromised by the research uses of the individual information. Levine suggested that it might "trivialize" the importance of informed consent if an individual were asked to simply check a box on a consent form, agreeing that the information could be used for all other legitimate research purposes.

MODELS FOR ENSURING DATA ACCESS AND PRIVACY PROTECTIONS

Myron Gutmann began by observing that the general issues surrounding data confidentiality are well reported in *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data* (National Research Council, 2007: Chapter 2). Data confidentiality concerns reflect principles for the protection of human subjects outlined in the Belmont report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), and they are supported in regulations by the Federal Policy for the Protection of Human Subjects (the Common Rule). There is an ethical consensus, Gutmann said, on the need both to protect human subjects and to share data.

In this general context, Gutmann said that protection of education data is special because Family Educational Rights and Privacy Act (FERPA) legislation and regulations are added on to the core protections of the Common Rule. The Common Rule was designed to protect human subjects in a research environment, assumes that prior informed consent is fundamental, and uses a "reasonable" standard for protection. In contrast, FERPA was designed to protect students and their families in an educational environment, does not assume prior informed consent, and has "a much more absolute standard for protection." Despite these differences in the law and regulations, he said, in reality, researchers do obtain informed consent for the use of school records, but they get this consent from school administrators, rather than from parents or students.

The Problem: With a Focus on Spatial Data

Gutmann illustrated the process of disclosure review (i.e., review of data sets to assess and prevent disclosure of individually identifiable information) using the analogy of trying to find Waldo in the children's book, *Where's Waldo?* (Handford, 1997). Gutmann said that, given five attributes about Waldo, he can easily find him in a small crowd. In reality, administrative and survey data sets are large; by analogy, locating Waldo based on five attributes is much more difficult when he is in a large crowd. However, spatial information or electronic monitors make it easier to find Waldo even in a large crowd, because all explicit geographic locations are identifiable (VanWey et al., 2005). For example, if one knows that Waldo always stands next to the ring toss at the carnival, or if Waldo always carries around a radio transponder that signals his location, he stands out in the crowd.

Gutmann explained that, when reviewing possible dissemination of data on a particular topic, it is important to recognize that these data may not be the only published source of information about that topic. For example, he might publish a map of Washtenaw County, Michigan, and list three attributes of an individual living in that county. This may pose little risk of revealing that individual's identity, because there are many individuals with those three attributes across the entire county. However, because there may be only one individual with these attributes who lives in one particular city block, publishing this geographic information would be likely to reveal the individual's identity. In the case of education records, if a few attributes of an individual student were published, along with that student's school, then the individual student would be easily identifiable.

Protecting Confidentiality: Goals and Options

Gutmann outlined two goals for protecting confidentiality when sharing or disseminating microdata: (1) to eliminate direct identifiers and (2) to eliminate unique individuals in small cells. Options for achieving these goals in tabular data with area identifiers but no precise spatial locations include the following:

- aggregating values (e.g., into five-year age groups instead of single years);
- top-coding (recoding values so that extreme values are combined with less extreme values);
- swapping data across spatial units; and
- paying careful attention to easily identified categories of data, especially geography, schools, and clustered samples.

Gutmann said that *Putting People on the Map* (National Research Council, 2007) differentiates between technical and institutional options for protecting confidentiality. Technical options include replacing real data with synthetic data for potentially identifying attributes and creating secure data analysis systems. Institutional options, which focus on individuals and organizations rather than on technology, include contracts and data enclaves. Institutional options vary, based on the perceived level of risk of disclosure; more complex options are more expensive and make research more difficult. For example, Gutmann said, the University of Michigan houses one of the nine research data centers operated by the Census Bureau. The center is expensive and difficult to use for investigators who live more than 30 miles away.

Gutmann presented a figure illustrating the gradient of levels of risk and levels of protection required for different types of data sets (see Figure 5-1). Simple data sets, with little risk of disclosure and little risk of harm if there were a disclosure, can be made publicly accessible on the Internet. If the data set is more complex and the risk of harm from disclosure is greater, then the data producer might require a formal data use agreement before providing access. Some data sets include information about illegal or undesirable behavior, which people voluntarily provide. Because disclosure of individual identities in such data sets would pose a great risk of harm, the data producer might require a strong data use agreement. This agreement would be likely to include restrictions on technology and tech-

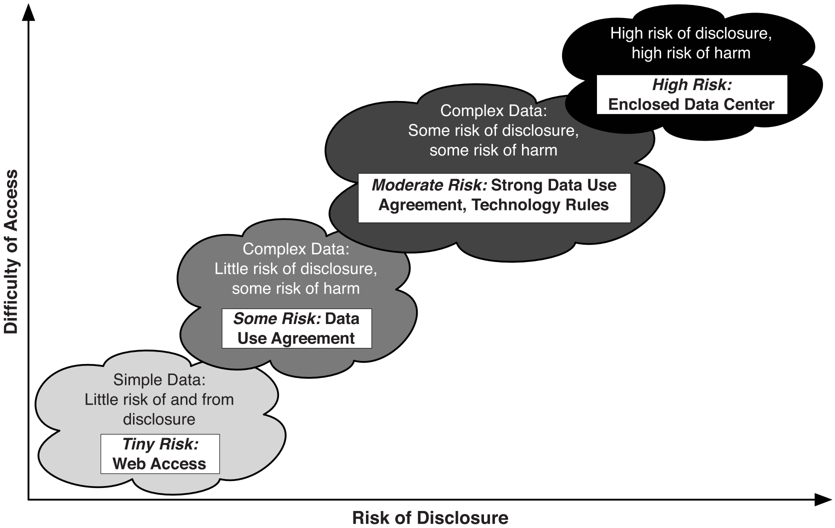

FIGURE 5-1 The gradient of risk and restriction.
SOURCE: Gutmann (2008).

nology access, similar to those included in National Center for Education Statistics licensing agreements (see Chapter 2). Finally, if there is a very high risk of disclosure and a high risk of harm, the data producer might "just lock things into an enclosed data center" and require the researcher to come and use the data onsite, Gutmann said.

Inter-University Consortium for Political and Social Research

Gutmann presented an overview of the activities of the Inter-University Consortium for Political and Social Research, the data archive he directs (see http://www.icpsr.umich.edu/). Experts at the archive conduct disclosure risk analysis of all data sets that enter the archive. In most cases, they remove all direct and indirect identifiers in the data sets and make them available through the Internet to users who have signed contracts with the consortium. Because most of these data sets are based on political opinion polls and information on voter behavior, Gutmann described their information as "really not very harmful." Other data, which are not useful for research without some indirect identifiers, are subject to further restrictions, ranging from easy to more stringent contracts. However, the consortium lacks authority to impose large fines on individuals who violate these contracts, in contrast to the National Center for Education Statistics (see Chapter 2). Finally, the most sensitive data sets, posing the greatest risk of harm to an individual whose identity might be disclosed, are housed in an onsite data enclave for use by researchers who appear in person.

The consortium houses and maintains data sets and conducts research with support from many organizations, including the National Institute of Child Health and Human Development and other agencies in the U.S. Department of Health and Human Services, the National Institute of Justice, and the consortium's member institutions. For example, the consortium is currently taking over dissemination of the National Longitudinal Study of Adolescent Health (Add Health) from the University of North Carolina at Chapel Hill. This study includes several different types of data, including survey (self-report) data and biomarker data. Biomarker data—indicators of disease or health, such as blood pressure, heart rate, and the presence or absence of certain molecules—include both physical specimens and digital representations. The study also includes analysis of ancillary data, including participants' high school transcripts. In compliance with FERPA, the investigators received individual written consent from each student in order to obtain access to the transcripts from their schools. To access these national data sets, a researcher must receive approval from the institutional review board at her or his home institution and provide a data security plan.

Future Directions in Reconciling Research Access with Confidentiality

In the future, Gutmann said, all administrative data sets should be online, with "reasonable" restrictions on access. These future data systems will be very distributed and have the capability to combine data "on the fly" to preserve confidentiality. They will have the capability to automatically recognize and solve confidentiality issues before the data reach the user's computer. In addition, these advanced future systems will build user communities on the basis of dynamic patterns of data usage.

Turning to future developments, Gutmann first contrasted FERPA with the Common Rule. He said that the Common Rule directs that requirements for reuse of data match the commitment contained in the original informed consent, although this does not always happen in reality. In contrast, because education records are administrative in origin, they are collected without informed consent, and FERPA rules absolutely forbid reuse of the data in ways that have any potential for identification of individuals, Gutmann said. In some cases, it is possible for researchers to use education records in compliance with FERPA by obtaining retrospective consent. For example, at the time they wanted to access high school transcripts for inclusion in the adolescent health study, researchers were already planning to go into the field to interview study participants. Near the end of the interview, they requested access to the students' transcripts, and over 80 percent gave consent. More often, however, researchers do not want to ask participants in a long-term study for consent, because they may decline and drop out of the study altogether.

Gutmann said that research organizations like his have found approaches that work well for archiving and using education data. He observed that researchers often think of school administrators as partners in the research process. Finally, Gutmann emphasized that current data protection schemes appear to work very well, with no known examples of individuals having been harmed by confidentiality breaches (see Chapter 1). This is important to keep in mind, he said, so that confidentiality does not become a higher priority than conducting good research.

In response to a question, Gutmann said he did not know of any studies about obtaining retrospective consent. Levine said that the process for obtaining a waiver of the informed consent requirement of the Common Rule from an institutional review board was well understood and used to facilitate research. Gutmann agreed, but cautioned that, in some cases, an institutional review board might grant a waiver under the Common Rule, but not under FERPA. Levine responded that FERPA was silent on this point. Because the law does not elaborate on the possibility of a waiver of informed consent, she said, lawyers and researchers can only assume that a waiver is not allowed. Gutmann responded that the Common Rule regulations have been revised several times over the years, on the basis of the experiences of the large community of data producers and researchers that the regulations govern. In contrast, he said that the Department of Education recently announced the first proposed revisions to FERPA regulations (see Chapter 2).

Barbara Schneider said that universities are careful to comply with FERPA and that she knows of many planned studies in which the researchers have decided to obtain written consent for access to education records rather than seek a waiver. Robert Boruch observed that research partnerships represent the future of social science research and that investigators would not be able to obtain more data without cooperating with education agencies and other data producers. He urged the research community to share experiences and approaches to obtaining access to data, including memoranda of understanding with education agencies and legal arguments.

IMPACT OF THE HIPAA PRIVACY RULE ON RESEARCH

Roberta Ness (University of Pittsburgh) shared a survey she conducted to assess the impacts on health research of the HIPAA Privacy Rule. The survey was commissioned by an Institute of Medicine (2008) committee as part of a larger study of the impacts of this law on health research.

Privacy Rule Protections

Ness explained that Congress enacted HIPAA in 1996 partly because the Common Rule did not definitively protect the privacy of individually identifiable health information. The law was designed, she said, to protect the privacy of medical records. The Privacy Rule implementing HIPAA (U.S. Department of Health and Human Services, 2000, 2002) permits health care provider organizations to disclose individually identifiable health information for research purposes only if the researcher has obtained written consent from each patient or, if that is impractical, a waiver of this requirement from an institutional review board. Although the rule does permit health care providers to disclose limited data sets with all identifiers removed[3] to researchers who sign a formal data use agreement, Ness said that these data sets cannot be linked to any other medical records and are not useful for research. The HIPAA Privacy Rule also permits disclosures to public health authorities without written consent for the purpose of public health surveillance. Institutional review

---

[3] The rule specifies 18 identifiers that must be removed, including geographic information and dates related to the individual.

boards at universities, hospitals, and other organizations implement the HIPAA Privacy Rule.

Ness said that an institutional review board will generally provide a waiver of the informed consent requirement if it determines that the proposed research presents a low risk of disclosure of individually identifiable information and that the proposed research could not be conducted without the waiver. She observed that there is "a great deal of local interpretation" of these two conditions.

Survey Content

Ness presented several news reports describing epidemiological and clinical research studies that were halted or slowed after the enactment of HIPAA. Describing the reports as "worrisome," she noted that they provide only anecdotal evidence about the possible effects of the HIPAA Privacy Rule on health research. The Institute of Medicine (IOM) commissioned a national survey in 2007 in order to make a more informed assessment of the effects of the rule, she said (Ness, 2007a).

The IOM survey was conducted in collaboration with 13 epidemiology professional societies. Each society contacted all of its active members, requesting that they respond to an anonymous, web-based survey about the HIPAA Privacy Rule. A total of 2,805 individuals accessed the website. Among this group, 1,527 indicated that they had submitted a research proposal to an institutional review board since the enactment of HIPAA, and the answers of this smaller group were analyzed.

Ness explained that the survey was designed to ask about positive and negative influences of the HIPAA Privacy Rule. First, the survey presented questions with quantitative responses, such as such as how frequently respondents collected various types of data, changes in the numbers of participants recruited before and after implementation of the Privacy Rule, and the level of difficulty encountered when seeking waivers or approval for release of deidentified data sets from the institutional review board. A second group of questions focused on respondents' perceptions of the ease and difficulty of conducting research under the HIPAA Privacy Rule and its impact on privacy and confidentiality, using a 5-point Likert scale. Third, the web-based survey presented five hypothetical research proposals, asking respondents whether their institutional review board would approve them. Finally, respondents were asked open-ended qualitative questions, including a request for stories about HIPAA.

Analysis of Responses

The respondents were predominantly women (59 percent), mostly employed in academia (66 percent), and they varied widely in age. In describing their perceptions, a large majority (84 percent) rated the degree to which the HIPAA Privacy Rule made research easier as low, at 1 to 2 on a 5-point Likert scale anchored at 1 = none. Responding to another question, a somewhat smaller majority (67.8 percent) rated the degree to which the Privacy Rule made research harder as high, at 4 to 5 on a 5-point Likert scale anchored at 5 = a great deal. Few respondents (10.5 percent) perceived the rule as having strengthened public trust a great deal, and only about one-quarter believed that the rule had greatly increased participant confidentiality. With respect to cost and delay, about half of the respondents perceived the rule as adding costs and delaying time to completion by a great deal.

Ness was surprised that more respondents indicated that the rule had a negative effect on the protection of human subjects than the number of respondents indicating that the rule had a positive effect. Ness said this response was "almost bizarre on the face of it, because, of course, this is legislation that was purposely enacted to improve the protection of human subjects." However, she went on to say that the respondents explained their responses in the qualitative section of the survey, indicating that they viewed the burden of paperwork resulting from adding HIPAA to the Common Rule as so great that medical patients no longer understood what they were giving informed consent for.

She said that about 15 percent of respondents indicated that, although their research proposals were approved by the institutional review board, the health care organization holding the medical records would not allow access because of Privacy Rule concerns. An additional 11.5 percent of epidemiologists surveyed had conceived of a study but not submitted it to an institutional review board because they thought they would be unable to obtain approval due to the Privacy Rule. More than half of the respondents said that an application they had submitted to an institutional review board was strongly adversely impacted by HIPAA.

Presenting the responses to the case study section of the survey, in which respondents were asked whether their institutional review board would approve five different types of studies, Ness said that the key finding was the wide variability (see Table 5-1). Such wide variation indicates that institutional review boards are interpreting the Privacy Rule in very different ways, Ness said.

In response to the final section of the survey, inviting HIPAA stories, Ness said they received almost 500 written responses, reflecting the "angst that's out there." A total of 90 percent of the stories were negative,

TABLE 5-1 Responses to Case Studies, Institute of Medicine Survey

	Would Your IRB Approve This Study?, Number (%)					
	No	Yes, Unconditional	Yes, with Waiver	Yes, with Approval	Yes, Other Conditions	Don't Know
Participants from medical records contacted for interview/blood draw	184 (12.56)	123 (8.5)	262 (18.0)	522 (35.9)	135 (9.3)	229 (15.7)
Participants from cancer registry contacted to consent for interview	196 (13.5)	157 (10.8)	261 (18.0)	468 (32.3)[a]	175 (12.1)	193 (13.3)
Tissue bank to supply deidentified data for assay not in original consent	222 (15.6)	199 (14.0)	159 (11.2)	262 (18.4)[b]	291 (20.4)[c]	290 (20.4)
Medical record review from subjects now dead	61 (4.7)	435 (33.8)	280 (21.8)	109 (8.5)[d]	121 (9.4)	281 (21.8)
Limited data set from another hospital; research cannot be done without some identifiers	239 (20.2)	58 (4.9)	260 (21.2)	427 (36.0)[e]	123 (10.4)	317 (26.7)

NOTE: IRB = institutional review board.

[a] With physician approval.

[b] With authorization and reconsent from patients.

[c] Limited data set or other special circumstances.

[d] With approval from executor of estate.

[e] With limited data set agreement.

SOURCE: Ness (2007). Copyright 2007 American Medical Association. Reprinted with permission.

5 percent were neutral, and 5 percent were positive. For example, one respondent wrote (Ness, 2007b):

> An already cumbersome patient consent form now has an additional page-and-a-half explaining HIPAA restrictions. This detracts from the informed consent process pertaining to the more critical issue: the actual medical risks and benefits of participating.

In general, the written responses indicate that the HIPAA Privacy Rule had not stopped health research, but it had slowed research progress and increased costs. Many respondents expressed the view that the Privacy Rule was hurting public health surveillance and causing confusion in the public health community. Ness reminded the group that the HIPAA Privacy Rule specifically permits disclosure of individual health information for public health surveillance.

Summarizing the IOM survey results, Ness reiterated that only about one-quarter of the respondents believed that the Privacy Rule had enhanced privacy. Among all respondents, the rule was seen as having a more negative than positive impact on the protection of human subjects. The analysis of the responses suggests that institutional review boards around the nation interpret the rule in quite different ways, making it unclear whether many of the problems described in the survey are a function of the Privacy Rule itself or local institutional review board interpretation of it. She said that the limitations of the survey include respondent bias; it may be that the respondents were those who feel most negatively about HIPAA. Another limitation is that it was not possible to calculate a response rate, because of the anonymous Internet process.

In discussion, Ness explained that the primary audience for the survey was the IOM committee, and that the survey reached a wider audience through publication in the *Journal of the American Medical Association* (Ness, 2007a). She said that representatives of two agencies in the Department of Health and Human Services—the Office for Civil Rights, which leads implementation of the Privacy Rule, and the National Institutes of Health, which funds a large amount of health research—had attended the IOM committee's meetings and were concerned about the wide variation in local interpretation of the rule. She observed that the IOM committee would decide how to respond to the problem in its final report.

Stephen Plank (Johns Hopkins University) asked if the word "surveillance" raised public fears. Ness responded that the IOM committee had commissioned a study of public attitudes toward privacy and health research (Westin, 2007), which found that language had a powerful influence on individuals' willingness to allow access to their health information.

A NEW APPROACH TO HEALTH DATA STEWARDSHIP

P. Jon White (Agency for Healthcare Research and Quality) opened his remarks with the observation that health care in the United States has a quality problem. The Institute of Medicine (2000) found that between 50,000 and 100,000 deaths were caused each year by medical errors, and more recently, McGlynn (2003) found that health care recipients got the recommended level and type of care only about 55 percent of the time. At the same time, health care has a cost problem. The current annual expenditure of $2.2 trillion represents a significant fraction of the nation's gross domestic product, and health care costs are rising steeply, at an annual average rate of 6.5 percent.

Current Efforts to Measure Health Care Quality

White said that one proposed solution to both of these problems is to pay for quality, rather than paying for individual visits to the doctor or for individual procedures or treatments. The Agency for Healthcare Research and Quality is one of several organizations working toward this solution; these organizations all face the key question of how to set the health care quality goals that would guide payments. In one effort to answer this question, his agency awarded grants to support health care information technology systems for "enabling quality measurements."[4] In addition, the Department of Health and Human Services' Center for Medicare and Medicaid Services funded six regional pilot projects to provide better quality information for Medicare beneficiaries.

White provided an example to illustrate a key issue in using administrative—or in this case, electronic claims—data to assess health care quality. If one wanted to assess the quality of Dr. White's treatment of diabetic patients, any single payer could provide claims data for only 10 to 15 percent of Dr. White's diabetic patients. Doctors have successfully argued that measures based on their treatment of these small samples of patients are inaccurate. To address this problem, the Agency for Healthcare Research and Quality and other organizations are beginning efforts to assemble health claims data from multiple payers, including the information assembled by the technology systems and regional pilot projects described above.

The agency has also funded development of 14 "chartered valued exchanges" around the country. These are coalitions of health care providers, payers, patients, and regulators who receive data from Medicare and from local payers and health care providers and try to use these data to

---

[4] See http://grants.nih.gov/grants/guide/rfa-files/RFA-HS-07-002.html.

measure quality. White explained that the Centers for Disease Control and Prevention's Biosense Health Surveillance Program[5] taps into the existing streams of data in hospital information systems around the country and "sends it up to the mother ship in Atlanta." Analysts there can monitor spikes in certain diseases, medical conditions, or symptoms. In addition to these federal efforts, health insurance companies and health maintenance organizations are working to develop measures of health care quality. The New York attorney general's office has signed agreements with several major health care organizations to rank doctors based on quality of care, rather than on how much they cost the organization (New York State Attorney General's Office, 2007). Other organizations that are trying to assemble and analyze health care data to develop measures of quality include the National Quality Forum, a public–private partnership, Google, and Microsoft.

All of these efforts face the question of who owns the health care data, White said. In the past, medical records were maintained in paper files, making it easier for any single doctor or hospital to own and keep them. With the change to digital records, it is possible for many individuals and organizations to own copies of health care records. White explained that he and his colleagues use the word "stewardship," which he defined as "taking care of something that doesn't belong to you." He has been engaged in discussions of health data stewardship with many organizations over the past few years, including the Ambulatory Quality Alliance, the National Committee on Vital and Health Statistics, which advises the Secretary of Health and Human Services, and the American Medical Informatics Association.

These initiatives also face privacy and security issues, as individual health records are protected by HIPAA, the Common Rule, and state and local laws and policies. In 2005, the Agency for Healthcare Research and Quality helped to fund a collaborative effort among more than 30 states and territories to study their privacy laws and regulations governing medical records. As a result of these studies, the participants have begun working to harmonize these laws and regulations, both within and across states.

A Data Stewardship Entity

Returning to the concept of stewardship, White explained that the idea of assembling multiple sources of data in order to improve health care quality emerged several years ago in the Ambulatory Quality Alliance. The alliance includes representatives of White's agency, two physi-

---

[5] See http://www.cdc.gov/BioSense/.

cians' organizations, and an association of health insurance companies. The alliance members recognize that doctors, laboratories, health insurance plans, and patients all have separate pieces of the health care information needed to measure quality. Through discussion, they developed principles for sharing and aggregation of these disparate sources of data, including (Ambulatory Quality Alliance, 2006):

- transparency with respect to framework, process, and rules;
- measurement of provider performance derived from standardized metrics and data collection protocols that can be compared with national, regional, or other suitable benchmarks and otherwise assists in the analysis of assessments of health care quality and cost of care;
- useful data for physicians to improve the quality and cost of care they provide to their patients and other appropriate purposes (e.g., maintenance of certification);
- public reporting to consumers of user-friendly, meaningful, and actionable information about physician quality and cost of care; and
- the collection of both public and private data so that physician performance can be assessed as comprehensively as possible.

White explained that, as the Ambulatory Quality Alliance members discussed these principles, they reached agreement on the need for a new health care data stewardship entity. When developing the mission and scope of the entity (Ambulatory Quality Alliance, 2006), they were unclear about whether the entity would simply set guidelines for assembling and managing data or would actually serve as a data archive. To solicit answers to this and other questions about the entity, the Agency for Healthcare Research and Quality published a request for information. Over 100 public and private health care organizations and individuals responded to the request, and the agency published a qualitative summary of their comments (Agency for Healthcare Research and Quality, 2007). The varied responses included significant support for both possible roles of the entity: setting guidelines for data stewardship and acting as the data steward. At the same time, some respondents expressed significant concerns, and some were completely opposed to the idea of sharing their personal medical records. White said he found it very valuable to hear and understand these views from the public.

Near the end of his presentation, White posed several questions that his agency and others are discussing as they consider the possibility of creating a health care stewardship entity:

- What is your problem? What are you trying to address? What do you need to do this for?
- Do you need a referee to address your problem by helping to set and enforce the rules of the game?
- Do you need someone to hold the information you need to address your problem?
- How do you avoid unintended consequences, including breaches of privacy and confidentiality?

White closed by warning that websites exist today, at which, for a small fee, an individual can enter a medical condition and receive a list of people who have that condition. The website managers gather the information from sources that are not governed by HIPAA. Finally, he said there are many questions and no conclusions.

In discussion, White observed that lobbyists on Capitol Hill are telling Congress now that the patient ought to control her or his health records, although the doctor and the insurance company should also be allowed to access the records. White said that such proposals miss the possibility of using medical records for research that could improve the health care system for the public good.

DISCUSSION: IMPLICATIONS FOR RESEARCH USING EDUCATION RECORDS

Reflecting on Lizanne DeStefano's earlier presentation about development of partnerships between researchers and schools (see Chapter 4), Miron Straf said that statistical agencies should work with administrative agencies, helping them to develop their data systems for statistical use. He asked whether there was a federal role in providing this type of assistance to education agencies.

Supporting Research Partnerships Through Trust and Technical Assistance

Marilyn Seastrom replied that 27 states currently have grants from the Department of Education's Institute of Education Sciences to develop longitudinal databases of education records and that more funding will be provided to the states in fiscal year 2009 (see Chapter 2). One requirement of these grants, she said, is that the states make the databases user-friendly and accessible to researchers. She said that, even if FERPA were changed to make data-sharing easier, states and school districts might still refuse researchers' requests for data access if they lack the resources and technical capacity to do so.

DeStefano responded that, in addition to trust, the motivation of research partners is also important. When a school or state education agency does not have a strong motivation to participate in a research project, she argued, the agency's leaders are more likely to say that they cannot share data because of FERPA. Seastrom agreed that trust and the development of relationships would continue to be very important for researchers to gain access to education data.

Paula Skedsvold (American Educational Research Association) asked whether legislative changes were needed in FERPA to clarify the meaning of research "for, or on behalf of" an education agency. She observed that, in response to the American Educational Research Association's survey about FERPA, some respondents indicated that they had simply abandoned research projects, because they could not obtain access to the education records they needed.

Seastrom asked workshop participants to describe how much state or local education agencies and researchers themselves alter education record data to protect confidentiality.

Schneider replied that, in one case, her team had helped to deidentify a file of teacher information. She observed that researchers outside her team who wish to use the file are required to apply to their own institution's institutional review board and to the state of Michigan's institutional review board for the use of data, providing a data protection plan along with other information about the proposed research.

DeStefano and others agreed that researchers should provide technical assistance to state and local education agencies to increase their capacity in techniques of deidentification.

Weighing Risks and Benefits of Disclosure and Research

Martin Orland proposed that Gutmann's matrix of risk and restriction (see Figure 5-1) should include another dimension—the likelihood of harm. Gutmann responded that the matrix included risk of harm and disclosure, and Orland replied that risk of harm and risk of disclosure should be two different dimensions. For example, he said, the risk of a nuclear power plant accident is minuscule, but this unlikely event could cause "enormous" harm. Orland expressed concern that the harm caused by even one disclosure of individually identifiable information could adversely affect the entire research environment, especially in light of the public and congressional concerns about privacy that had generated the HIPAA legislation.

Gutmann agreed with Orland that it is important to differentiate between the risk of disclosure and the risk of harm. For example, he said, there is almost nothing on the short-form census questionnaire that should cause an individual to be concerned if it were publicly revealed. In contrast, he said, other data that individuals provide in surveys or that are included about them in administrative records would pose great risk of harm if they were revealed, as shown in his graphic (see Figure 5-1). He went on to explain that he is much more worried about the potential harm disclosure could cause to groups of individuals than he is about the harm to researchers of limited access to data. He said that the people who fund his data consortium do so because they do not want to see a front page story in the news about any revelation of individual identities based on data they collect.

Gerald Gates added that, unlike Gutmann, privacy laws do not distinguish between more sensitive and less sensitive individual information; these laws simply state that individual information cannot be disclosed. Although institutional review boards in federal agencies consider the sensitivity of different data sets when determining how to protect them, they focus primarily on complying with the letter of the law by protecting against any disclosure of any individually identifiable information. Levine responded that one criticism of institutional review boards is that, when reviewing a research proposal, they fail to distinguish between the risk of disclosure and the magnitude of harm that a disclosure would cause.

Gutmann responded that institutional review boards do not always take advantage of the flexibility they have to allocate their time and resources. For example, the University of Michigan institutional review board has explicit rules stating that research proposals to use data from a list of specific deidentified public data sources (including the Census Bureau, his institute, and other sources) do not require institutional review board approval (University of Michigan, 2008). Therefore, the board does not need to devote resources to reviewing these research proposals and can focus on other proposals in which protecting human subjects is more important. Gutmann suggested that the research community continue to work with institutional review boards to make sure that they are devoting their resources where they are most needed, especially because he sees most institutional review boards as "overwhelmed."

Seastrom responded that, while she agreed with Gutmann, the exact opposite would be the case for a disclosure review board. This type of board would be very concerned about what type of public information a researcher would add to a data set, she said, and Gutmann agreed.[6]

Straf said that it was important not only to distinguish between the risk of a potential disclosure and the harm that could be caused, but also

---

[6] The University of Michigan (2008) policy states that a researcher who plans to merge more than one public data set and recognizes that this may increase the risk of identification of individual research participants should consult the institutional review board.

between the risk of disclosure and the benefits of research. Gutmann responded that earlier workshop sessions had illustrated the benefits of using education records for research (see Chapter 3). DeStefano said that, in the partnership model, education agencies and researchers discuss the specific benefits of particular research projects, rather than considering the general benefits of research to society; she observed that the University of Illinois institutional review board had made note of these specific benefits when reviewing research partnership proposals.

Ness said that the Institute of Medicine committee extensively discussed the risks and benefits of research using individual health information. The committee commissioned surveys showing that the public is "hungry" for health information, and thinks that the United States should remain the world leader in generating new medical knowledge (Westin, 2007). She suggested placing the new knowledge resulting from research in "a very central position" when weighing research benefits and privacy risks. Levine agreed that the public increasingly recognizes the importance of health as a public good, saying that the public should view education in the same way.

Boruch said that an early report on privacy and confidentiality by the Committee on National Statistics (National Research Council, 1979) included an analysis of how people react to a request for personal information presented in different ways. He suggested that the survey of public attitudes commissioned by the Institute of Medicine Committee might be a valuable resource for understanding how to frame such requests, which is a challenging task across fields of social science research; Ness said the survey is publicly available (Westin, 2007).

Schneider urged the American Educational Research Association to continue providing professional development on keeping data confidential. While acknowledging her fear that a breach of individual identity was inevitable, she said it was critical to educate the research community about confidentiality and how best to safeguard it.

6

Reflections and Next Steps

PLANNING COMMITTEE REFLECTIONS

In the final session, members of the workshop planning committee reflected on what they had learned about reconciling access, privacy, and confidentiality of education records and offered suggestions for future research and policy. Felice Levine noted that, on one hand, the workshop had described several different models that allow researchers to access education and other administrative records while protecting confidentiality and had also illuminated the benefits of research using education record data. On the other hand, she said, the workshop discussions had clarified the challenges that result from the failure of the Family Educational Rights and Privacy Act (FERPA) to reconcile privacy with research access in an effective way.

Robert Boruch reiterated his earlier call for researchers to use more uniform language to communicate with the public and clarify the distinction between statistical and administrative uses of administrative data. He said that the discussions about models of research access were very helpful, asking researchers and public agencies to share their formal data-sharing agreements and memoranda of understanding. Boruch argued that sharing these agreements is essential to alleviate fears about complying with FERPA and other privacy laws. Boruch said that changes to the FERPA law or regulations, such as redefining "educational institution" to include state education agencies, as called for by Steven Winnick (see Chapter 2), would help to facilitate researchers' access to data. And he said that the process of obtaining a waiver of the FERPA informed consent requirement from a state or local education agency presents an opportunity to forge a research partnership.

Thomas Plewes (National Research Council) observed that federal agencies have developed innovative approaches to providing access to data for research purposes, including research data centers, data enclaves, and the data licensing agreements pioneered by the National Center for Education Statistics. However, these innovations at the federal level have not yet been tried by state or local education agencies. He noted that the No Child Left Behind Act generates the need to gather data on student performance and also drives the need for more education research. Constance Citro (National Research Council) agreed that the workshop had highlighted the value of different models of access tailored to different types of data, such as the Census Bureau's research data centers, which severely limit access, as is appropriate for the sensitive data they maintain. Noting that state and local education agencies are most affected by FERPA, Citro asked which of the federal models would be most effective in helping these agencies provide data access while protecting confidentiality.

Martin Orland said he had learned that "federalism is alive and well" through the workshop discussions, highlighting the question of what the federal role should be. He noted that Congress had not anticipated the possibility that education records might be used for research purposes when it wrote FERPA. He called for changes in the law that would recognize the value of using school records to benefit research and improve education policies and practices. He cautioned against concluding that "all is well," simply because presenters had described a few successful models of research access. In addition to changing FERPA, he said, success in using education record data for research purposes requires four critical conditions:

1. There is researchable data.
2. There is a confluence of interest between a researcher and an education agency.
3. Time and commitment are available to build trust between the researcher and the agency.
4. The education agency has the technical capacity to share data while protecting confidentiality of individual information.

Helen Ladd said that the critical issues of research access and confidentiality protection revolve around the use of state education data, reflecting the reality that education is a state function under the U.S. Constitution. She called for increased clarity in the Department of Education's guidance about FERPA, noting that state and local education agencies' nervousness about data-sharing agreements poses a barrier to good research. Ladd agreed with Orland that FERPA is not the only barrier to high-quality education research, observing that community colleges in North Carolina have information technology systems that are not capable of providing deidentified data to researchers. In addition, she reminded participants that Florida's education agency, with one of the most developed education data systems in the nation, lacks staff to respond to the many requests for access from individual researchers. Finally, she said that both researchers and education agencies must continually work to build trust.

Barbara Schneider noted that many speakers had agreed on the need to fix FERPA, because of its "chilling effect" on education and public health research. At the same time, she said that the workshop discussions had clarified the tensions between the benefits of research access and the harm that could come to an individual from an "inevitable" breach of sensitive personal information. Schneider repeated her earlier call for professional development about privacy and confidentiality in the education research community.

REFLECTIONS BY KENNETH PREWITT

The workshop planning committee invited Kenneth Prewitt, a former Census Bureau director and long-time member of the Committee on National Statistics, to provide concluding reflections at the end of the workshop.

Administrative Data and Survey Data

Prewitt said that researchers must use administrative data because they have so much to offer. He said that, although many social science researchers are not very familiar with data mining, this technique for analyzing administrative data represents a more important methodological breakthrough than any advance in survey methodology. For example, the Bureau of Economic Analysis now purchases credit card data from banks, mining the data for information to incorporate into its economic modeling.

Administrative and survey data differ in several important ways, he said. Administrative records are "theory indifferent," poor in number of variables, and rich in number of cases. Because they are gathered for administrative purposes, these data are not based on any theory of human behavior and include only a very limited number of variables, which are related to the administrative purposes. For example, an agency may gather data on an individual's earnings for purposes of providing

TABLE 6-1 Administrative and Survey Data

Characteristic	Administrative Records	Survey Data
Theory	Indifferent	Embedded
Number of variables	Poor	Rich
Number of cases	Rich	Poor

SOURCE: Presentation by Kenneth Prewitt (2008).

social benefits, without gathering data on any other variables related to that individual. This leads to creation of very large data sets that are thin in the number of variables and are not designed to test any particular theory. In contrast, survey data are "theory embedded"—that is, the design of the survey and therefore the characteristics of the resulting data are informed by theory—and rich in variables. Survey data are poor in number of cases, because of the expense of administering a survey; for example, a major national survey of public health includes only 5,000 cases (see Table 6-1).

These characteristics affect privacy and confidentiality, Prewitt explained. Because administrative data sets are thin in variables, researchers must often link them to other data sets in order to address important research questions. But this linkage magnifies the risk that individual information could be identified.

The two data sources also vary in quality, Prewitt said. The accuracy of variables in administrative data depends in part on the goals and missions of agencies that gather these data. For example, the Social Security Administration is unlikely to tolerate errors in age data, while the Department of Housing and Urban Development may be less concerned about accuracy in this variable. Prewitt noted that survey theory and practice are well developed, with a history dating back to the 1930s, and today students can take many classes to learn about how to gather high-quality survey data. Asking how many classes are available on the quality of administrative data, Prewitt called for increased study and discussion focusing on the quality of these very important new data sources.

Developing Research Partnerships

Prewitt observed that, at the federal level, relationships between researchers and data-producing agencies have matured over the past 15 years. Their ongoing discussions of research access, privacy, and confidentiality have informed the development of new models that reconcile access with protections, such as data enclaves and data licensing agreements. Prewitt suggested that a similar dialogue was needed at the state level, in order to support development of similar models there. He encouraged researchers to try to understand the motivations and ethical and emotional concerns of state officials as part of that dialogue.

Echoing Boruch, Prewitt also called for increased sharing of formal memoranda of understanding between researchers and education agencies, in order to address state and local officials' fears about sharing data while complying with FERPA and other privacy laws. When considering concerns about privacy and data confidentiality, he said, it is important to distinguish between the harm to an individual whose data might be released to the public and the harm that could result to the research enterprise. He said that he had participated in a discussion earlier that day about whether and how the National Center for Health Statistics might release DNA data. A knowledgeable expert said that, if individual DNA data ever were inadvertently released and this became publicly known, this would effectively mean the end of the National Center for Health Statistics. Prewitt urged the research community to assume responsibility for protecting data confidentiality, along with the data providers, in order to avoid harm to both research and agency missions.

Prewitt called for greater clarity about the different uses of administrative data. These data are used for administrative purposes, for research purposes, and also for program evaluation and policy design, he said. For example, the No Child Left Behind Act requires states to gather data for program evaluation. He asked the research community to avoid a tendency to project the way it wants to use administrative data onto all of these other uses.

Next Steps

Prewitt said that the workshop had been valuable in developing new ideas related to access to school record data for use in education research, going beyond a narrow focus on changing the legislative language in FERPA. He urged participants to offer comments to the Department of Education on its proposed revisions to the FERPA regulations. Based on his experience at the Census Bureau, he said, "I can tell you that comments are read carefully," as agencies try to be responsive to interested stakeholders.

He noted that similar conversations about balancing data access for research and confidentiality were going on in health and other sectors, and he urged continued discussion across sectors. Such conversations about the complexities and challenges—both across sectors and across levels of government—will help to make agencies and researchers more comfortable with sharing data, he said. Finally, he said he thinks of the current situation as involving two sides. On one side are the school system, parents, and grandparents wishing for a better public school system. On the other side is a large education research enterprise, including the 25,000 experts who participate each year in the American Educational Research Association's annual conference. Somewhere in between the analytic capacity of the researchers and the concerned students and families are the rich education data sets. The workshop discussions, he said, were about the way in which this research capacity can address the social challenge of improving education by using the rich data sets to generate evidence.

FINAL DISCUSSION

Michael Feuer (National Research Council) responded to Prewitt's remarks about the quality of administrative data, focusing particularly on educational test score data. When thinking about linking test data with other data sets, Feuer said, it is important to consider how the validity of these test data might be compromised by the linkage. In addition, he cautioned that students' responses to certain conditions—such as high stakes attached to certain tests—may compromise their performance, affecting the quality of the resulting score data.

Moving on to FERPA, Feuer warned researchers to avoid exaggerating their complaints about privacy regulations, because this appears self-serving. He suggested considering other stakeholders' views of privacy and access issues, including the views of legal experts. Felice Levine responded that, as currently administered, FERPA does not address responsible research access to the data and is "almost exclusionary."

Levine responded to Prewitt's comments about the different uses of administrative data by observing that, while there has been more attention to these different uses for large administrative data sets, the federal government has provided little guidance. For example, the Common Rule provides guidance on how institutional review boards should review research proposals, but it devotes far less attention to how to protect privacy and confidentiality as the research plan is executed and when the results are being disseminated. In contrast, she said, she attended a recent meeting on biosecurity at which the entire discussion focused on dissemination and on the potential for unwanted uses of information. For example, if a researcher wants to publish the invention of an aerosol spray that could potentially be used to disperse airborne spores that could annihilate populations, biosecurity rules provide guidance on how to present the research findings in a way that would avoid this unwanted use. Levine then invited a team of researchers and public school representatives from Baltimore, Maryland, who participated in the workshop and have formed a research partnership, to offer any comments about the workshop or pose any final questions.

Ike Diibor (Baltimore City Public Schools) responded that the workshop had been very helpful in increasing his understanding of current activities in the research community. He said that he and other school officials have interpreted FERPA with the guidance of the school system's legal counsel, without devoting much thought to researchers' perspective on the law. Observing that the team was in the process of creating a formal research consortium, he said that the information gathered at the workshop about confidentiality and privacy would be very valuable.

Stephen Plank said that the workshop participants had seen the team of researchers and school officials from Baltimore try to practice the process of building trust and developing shared understandings. Reflecting on earlier discussions, he agreed with Helen Ladd that researcher–school system partnerships should not depend only on trust and personal relationships but should be codified in formal memoranda of understanding.

References

Abowd, J.M., Stephens, B.E., Vilhuber, L., Andersson, K.L., McKinney, M.R., Roemer, M., and Woodcock, S. (2005). *The LEHD infrastructure file and the creation of the quarterly workforce indicators*. Technical Paper #TP-2006-01. Suitland, MD: U.S. Census Bureau. Available: http://lehd.did.census.gov/led/library/techpapers/tp-2006-01.pdf [accessed July 2008].

Agency for Healthcare Research and Quality. (2007). *Summary of responses*. National Health Data Stewardship Entity Request for Information issued by the Agency for Healthcare Research and Quality, June 4, 2007. Rockville, MD: Author. Available: http://healthit.ahrq.gov/portal/server.pt/gateway/PTARGS_0_3882_816199_0_0_18/National-HealthDataStewardship_SummaryResponses.doc [accessed July 2008].

Ambulatory Quality Alliance. (2006). *Data sharing and aggregation work group*. Washington, DC: Author. Available: http://www.aqaalliance.org/datawg.htm [accessed July 2008].

American Educational Research Association. (2008). Letter on FERPA. *Educational Researcher, 37*(5), 304-306. Available: http://www.aera.net/uploadedFiles/Publications/Journals/Educational_Researcher/3705/07EdR08_302-306.pdf [accessed August 2008].

Bettinger, E.P., and Long, B.T. (2008). Addressing the needs of under-prepared students in higher education: Does college remediation work? *Journal of Human Resources.* Available: http://gseacademic.harvard.edu/~longbr/Bettinger_Long_-_Addressing_needs_of_under-prepared_students_(4-08).pdf [accessed October 2008].

Bifulco, R., and Ladd, H.F. (2007). School choice, residential segregation and test-score gaps: Evidence from North Carolina's Charter School Program. *Journal of Policy Analysis and Management*, *26*(1), 31-56.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2005). The draw of home: How teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis and Management*, *24*(1), 113-132.

Boyd, D., Lankford, S.H., Loeb, J.S., Rockoff, J., and Wyckoff, J. (2008). *The narrowing gap in New York City teacher qualifications and its implications for student achievement in high poverty schools*. Working Paper 2007-001. Stanford, CA: Stanford University Institute for Research on Education Policy and Practice. Available: http://irepp.stanford.edu/publications/working_papers/ [accessed July 2008].

Bradburn, N.M., and Straf, M.L. (2003). The eleventh Morris Hansen lecture. Information and statistical data: A distinction with a difference. *Journal of Official Statistics, 19(*4), 321-331.

Bryk, A.S., and Schneider, B. (2002). *Trust in schools: A core resource for improvement.* New York: Russell Sage Foundation.

Calcagno, J.C., and Long, B.T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance.* Working Paper. New York: Columbia University, Teacher's College, National Center for Postsecondary Research. Available: http://www.postsecondaryresearch.org/i/a/document/6964_CalcagnoLong.pdf [accessed July 2008].

Cook, P., MacCoun, R., Muschkin, C.G., and Vigdor, J. (2007). The negative impact of starting middle school in sixth grade. *Journal of Policy Analysis and Management, 27*(1), 104-121.

Cornell University. (2008). *VirtualRDC News@Cornell: What is the VirtualRDC?* Available: http://www.vrdc.cornell.edu/news/?page_id=9 [accessed May 2008].

Council of Chief State School Officers and Standard & Poor's. (2008). *SchoolDataDirect.* Available: http://www.schooldatadirect.org/ [accessed May 2008].

Ewell, P., and Boeke, M. (2006). *Tracking student progression: The state of the states*. Unpublished manuscript. Available: http://www.nchems.org/c2sp/sur/SURSurveyReport.pdf [accessed July 2008].

Fanning, J.P. (2007). *Policies and best practices for ensuring statistical and research confidentiality*. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services.

Gates, G. (2008). *Providing researchers with authorized, safe, useful access to administrative data*. Paper prepared for the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

Gutmann, M. (2008). *Protecting confidentiality in research data: Successful approaches in the academic research setting*. Presentation to the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

Handford, M. (1997). *Where's Waldo?* Somerville, MA: Candlewick Press.

Hanushek, E.A. (2007). *Plain talk with Eric Hanushek*. Washington, DC: Center for Analysis of Longitudinal Data in Educational Research. Available: http://www.caldercenter.org/news/ptEHanushek.cfm [accessed July 2008].

Hanushek, E.A., Kain, J.F., and Rivkin, S.G. (2004). Why public schools lose teachers. *Journal of Human Resources, 39*(2), 326-354. Available: http://www.utdallas.edu/research/tsp/pdfpapers/paper25.pdf [accessed July 2008].

Institute of Medicine. (2000). *To err is human: Building a safer health system.* Committee on Quality of Health Care in America, L.T. Kohn, J.M. Corrigan, and M.S. Donaldson, eds. Washington, DC: National Academy Press. Available: http://www.nap.edu/catalog.php?record_id=9728 [accessed November 2008].

Institute of Medicine. (2008). *Health research and the privacy of health information: The HIPAA privacy rule*. Available: http://www.iom.edu/CMS/3740/43729.aspx [accessed July 2008].

Jacob, B., and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226-244.

Jacob, B., and Lefgren, L. (2007). *The effect of grade retention in high school completion*. NBER Working Paper No. 13514. Cambridge, MA: National Bureau of Economic Research. Available: http://www.nber.org/papers/w13514 [accessed July 2008].

Kain, J. (2000). *Description of the UTD Texas Schools Project*. Dallas: University of Texas-Dallas. Available: http://www.utdallas.edu/research/tsp/pdfpapers/paper16.pdf [accessed July 2008].

Kain, J.F., and Singleton, K. (1996). Equality of educational opportunity revisited. *New England Economic Review*, May/June, 88-110.

Keesler, V., Wyse, A., Jones, N., and Schneider, B. (2008). *Calculating the ability of within-school teacher supply to meet the demands of new requirements: The example of the Michigan Merit Curriculum* (REL Technical Brief, REL 2008-No.005). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Available: http://ies.ed.gov/ncee/edlabs/ [accessed August 2008].

Lee, C., and Goldfarb, Z.A. (2006). Stolen VA laptop and hard drive recovered. *The Washington Post,* June 30, p. A-1. Available: http://www.washingtonpost.com/wp-dyn/content/article/2006/06/29/AR2006062900352.html [accessed July 2008].

Loeb, S. (2008). *University research and the Family Education and Privacy Rights Act*. Paper prepared for the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

McDonald, S., Andal, J., Brown, K., and Schneider, B. (2007). *Getting the evidence for evidence-based initiatives: How the Midwest states use data systems to improve education processes and outcomes*. Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Available: http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=29&productID=15 [accessed October 2008].

McGlynn, E.A. (2003). The quality of health care delivered to adults in the United States. *New England Journal of Medicine, 348*, 26.

Miranda, M.L., Kim, D., Overstreet, A., Paul, C.J., Hull, A.P., and Morgan, S.P. (2007). The relationship between early childhood blood lead levels and performance on end-of-grade tests. *Environmental Health Perspectives, 115*(8), 1242-1247.

Muschkin, C., and Ladd, H. (2008). *Research access to state education administrative data: The North Carolina Education Research Data Center*. Paper prepared for the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

National Center for Educational Achievement. (2008). *Data quality campaign*. Available: http://www.dataqualitycampaign.org/ [accessed June 2008].

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research.* Bethesda, MD: National Institutes of Health Office of Human Subjects Research. Available: http://ohsr.od.nih.gov/guidelines/belmont.html [accessed July 2008].

National Research Council. (1979). *Privacy and confidentiality as factors in survey response.* Washington, DC: National Academy Press.

National Research Council. (1993). *Private lives and public policies: Confidentiality and accessibility of government statistics*. Panel on Confidentiality and Data Access, G.T. Duncan, T.B. Jabine, and V.A. de Wolf, eds. Washington, DC: National Academy Press.

National Research Council. (2007). *Putting people on the map: Protecting confidentiality with linked social-spatial data.* Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. Washington, DC: The National Academies Press.

Ness, R.B. (2007a). Influence of the HIPAA privacy rule on health research. *Journal of the American Medical Association, 298*(18), 2164-2170. Available: http://jama.ama-assn.org/cgi/content/full/298/18/2164 [accessed July 2008].

Ness, R.B. (2007b). *Impact of implementation of the HIPAA privacy rule on research*. Presentation to the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

New York State Attorney General's Office. (2007). *Attorney General Cuomo announces doctor ranking agreement with GHI and HIP.* Albany: Author. Available: http://www.oag.state.ny.us/press/2007/nov/nov20b_07.html/ [accessed October 2008].

Prince, D., and Jenkins, D. (2005). *Building pathways to success for low-skill adult students: Lessons for community college policy and practice from a statewide longitudinal tracking study*. Olympia: Washington State Board for Community and Technical Colleges. Available: http://www.sbctc.ctc.edu/docs/education/ford_bridges/bldg_pathways_to_success_for_low-skilled_adult_stdts.pdf [accessed July 2008].

Privacy Protection Study Commission. (1977). *Personal privacy in an information age.* Washington, DC. Available: http://epic.org/privacy/ppsc1977report/ [accessed July 2008].

Purnell, R., and Blank, S. (2004). *Support success: Services that may help low-income students succeed in community college*. New York: MDRC. Available: http://www.mdrc.org/publications/399/abstract.html [accessed July 2008].

Ropp, M., Gosa, K., Schneider, B., McDonald, S.-K., and Dawson, M. (2008). *Creating high-impact information from longitudinal data*. Presentation to 2008 MIS Conference Longitudinal Data Systems Strand, February, San Francisco, CA. Available: http://nces.ed.gov/programs/slds/mis_08.asp#sessionV [accessed August 2008].

Schneider, B. (2008). *Lessons from the data: A collaboration between MDE/CEPI and MSU.* Presentation to the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

Sellers, J. (2008). *Florida's PK-20 education data systems.* Presentation to the Workshop on Protecting Student Records and Facilitating Education Research, April 24-25, National Research Council, Washington, DC.

University of Michigan. (2008). *Research using publicly-available data sets: UM policy*. Available: http://www.research.umich.edu/hrpp/Documents/datasets.html [accessed July 2008].

U.S. Department of Education. (2008a). Family educational rights and privacy: Proposed rule. *Federal Register, 73(*57), 15573-15602. Available: http://www.ed.gov/legislation/FedRegister/proprule/2008-1/032408a.html [accessed June 2008].

U.S. Department of Education. (2008b). *The ED*Facts *initiative*. Available: http://www.ed.gov/about/inits/ed/edfacts/index.html [accessed June 2008].

U.S. Department of Health and Human Services. (2000). Standards for privacy of individually identifiable health information: Final rule. *Federal Register, 65*(250), 82461-82510. Available: http://aspe.hhs.gov/ADMNSIMP/final/PvcPre01.htm [accessed October 2008].

U.S. Department of Health and Human Services. (2002). Standards for privacy of individually identifiable health information: Final rule. *Federal Register, 67*(157), 58181-53273. Available: http://www.hhs.gov/ocr/hipaa/privruletxt.txt [accessed December 2008].

VanWey, L.K., Rindfuss, R.R., Gutmann, M.P., Entwisle, B. and Balk, D.L. (2005). Confidentiality and spatially explicit data: Concerns and challenges. *Proceedings of the National Academy of Sciences, 102*(43), 15227-15342.

Westin, A. (2007). *How the public views privacy and health research*. Results of a national survey commissioned by the Institute of Medicine's Committee on Health Research and the Privacy of Health Information. Washington, DC: Institute of Medicine. Available: http://www.iom.edu/CMS/3740/43729/48528.aspx [accessed July 2008].

Appendix A

Workshop Agenda and Participants

AGENDA

Workshop on Protecting Student Records and Facilitating Education Research

April 24-25, 2008

April 24, 2008

8:30-8:45 AM Welcome and Introductions
Felice Levine, Chair

8:45-9:30 AM Session 1: Overview of the Issues
Miron Straf, Deputy Director, DBASSE, National Academy of Sciences

9:30-10:45 AM Session 2: Review, History, and Proposed Changes to FERPA Affecting Research Access
Moderator: *Robert Boruch*
- Ellen Campbell, Deputy Director, Family Policy Compliance Office, U.S. Department of Education
- Steven Winnick, Nelson Mullins Riley & Scarborough

10:45-11:00 AM Break

11:00 AM - 12:15 PM	Session 3: Reconciling Privacy, Confidentiality, Consent, and Access in Federal Statistical and Health Programs Moderator: *Constance Citro* • Gerald Gates, U.S. Bureau of the Census (Retired) • Jonathon White, Health IT Director, Agency for Healthcare Research and Quality
12:15-1:30 PM	Lunch
1:30-2:45 PM	Session 4: Reconciling Privacy, Confidentiality, Consent and Access: Department of Education Programs Moderator: *Constance Citro* • Ross Santy, Deputy Assistant Secretary for Data and Information, U.S. Department of Education • Marilyn Seastrom, National Center for Education Statistics, U.S. Department of Education
2:45-3:00 PM	Break
3:00-4:30 PM	Session 5: Options for Research Access to State Administrative Data That Protect Privacy and Confidentiality Moderator: *Martin Orland* • Jay Pfeiffer, Deputy Commissioner, Florida Department of Education • Helen Ladd, Stanford Institute of Public Policy, Duke University • Barbara Schneider, Michigan State University
4:30-5:00 PM	Wrap-Up and Issues of the Day *Felice Levine*, Chair

April 25, 2008

8:30-10:00 AM	Session 6: Critical Importance of Research from Administrative Record Systems Using Individually Identifiable Student and School Record Data Moderator: *Helen Ladd*

- Jane Hannaway, Director, Education Policy Center, Urban Institute
- Thomas Bailey, Director, Community College Research Center, Teachers College, Columbia University
- Susanna Loeb, Stanford University

10:00-10:15 AM Break

10:15 AM-12:00 PM Session 7: Other Models for Assuring Data Access and Privacy
Moderator: *Miron Straf*

- Myron Gutmann, Director, Inter-University Consortium for Political and Social Research, University of Michigan
- Lizanne DeStefano, Bureau of Educational Research, College of Education, University of Illinois at Urbana-Champaign
- Roberta Ness, Graduate School of Public Health, University of Pittsburgh

12:00-1:15 PM Lunch

1:15-3:00 PM Session 8: Feasibility of Change: Strategic Steps
Moderator: *Felice Levine*

- Kenneth Prewitt, Columbia University

Open Discussion

3:00 PM Adjourn

PARTICIPANTS

Thomas R. Bailey, Columbia University
Robert Boruch, Wharton School, University of Pennsylvania
Ellen Campbell, Family Policy Compliance Office, U.S. Department of Education
Jeanetta Churchill, Family League of Baltimore City
Constance Citro, National Research Council
Stephen Cohen, National Science Foundation
Elizabeth Cranley, University of Wisconsin-Madison
Matthew Dawson, Learning Point Associates
Lizanne DeStefano, University of Illinois
Ike Diibor, Baltimore City Public Schools
Bridget Dooling, Office of Management and Budget
Rachel Durham, Johns Hopkins University
Michael J. Feuer, National Research Council
Stefanie Fricano, California Legislative Analyst's Office
Norma Garza, U.S. Department of Education
Gerald Gates, U.S. Census Bureau (retired)
Aimee Guidera, Data Quality Campaign
Myron Gutmann, Inter-University Consortium for Political and Social Research, University of Michigan
Jane Hannaway, Urban Institute
Kimberly Harris, North Carolina State University
Jeanine Hildreth, Baltimore City Public Schools
Margaret Hilton, National Research Council
Dominique Johnson, Temple University
Helen Ladd, Duke University
Felice Levine, American Educational Research Association
Thomas Lindsley, National Center for Educational Achievement
Susanna Loeb, Stanford University
Mary Lummus-Robinson, National Center for Educational Achievement
Shelly Martinez, U.S. Office of Management and Budget
Roberta Ness, University of Pittsburgh
Denis Newman, Empirical Education, Inc.
Martin Orland, WestEd
Stephen Plank, Johns Hopkins University
Thomas Plewes, National Research Council
Kenneth Prewitt, Columbia University
Jeffery Rodamar, U.S. Department of Education
Margaret Ropp, Michigan Center for Educational Performance & Information
Michael Ross, National Center for Education Statistics, U.S. Department of Education

Ross Santy, Office of Planning, Evaluation, and Policy Development, U.S. Department of Education
Barbara Schneider, Michigan State University
Marilyn Seastrom, National Center for Education Statistics, U.S. Department of Education
Jeff Sellers, Florida Department of Education
Michael J. Siri, National Research Council
Paula Skedsvold, American Educational Research Association
Gerald Sroufe, American Educational Research Association
Miron Straf, National Research Council
Francie Streich, Spencer Foundation
Karen Studwell, American Psychological Association
Bridget Thomas, American Educational Research Association
Kevon Tucker-Seeley, Education Development Center, Inc.
Andrew White, National Center for Education Statistics, U.S. Department of Education
Dennis White, George Washington University
P. Jonathon White, Agency for Healthcare Research and Quality
Steven Winnick, Nelson Mullins Riley & Scarborough

Appendix B

Biographical Sketches of Planning Committee Members

Felice J. Levine *(Chair)* is executive director of the American Educational Research Association. Previously, she was executive officer of the American Sociological Association from 1991 to 2002. Her research specialties include children and youth and the dynamics underlying their social development. She also has done considerable work on research and science policy issues, academic and scientific professions, and the ethics of research. She is a member of the executive committee of the Consortium of Social Science Associations and served as chair from 1997 to 2000. She serves on the advisory committee of the National Consortium on Violence Research and the Research Advisory Committee for the American Bar Foundation. She also chairs the Social and Behavioral Sciences Working Group on Human Research Protections and served in 2001-2002 on the National Human Research Protections Advisory Committee. She is a fellow of the American Association for the Advancement of Science and the American Psychological Society and is a past president of the Law and Society Association. She has an A.B. in sociology and A.M. and Ph.D. degrees in psychology (social), all from the University of Chicago.

Robert F. Boruch is university trustee chair professor of education and professor of statistics at the Wharton School of the University of Pennsylvania. He is codirector of the Center for Research and Evaluation of Social Policy and codirector of the Policy Research, Evaluation, and Measurement Program, both in the Graduate School of Education. He has served on advisory committees for the U.S. Department of Education, the

National Institutes of Health, and many other federal agencies. He is also on the advisory boards for the Coalition for Evidence-Based Policy and the Society for Research on Educational Effectiveness and serves on the editorial board of *Evaluation Review* and other journals. He is an elected fellow of the American Academy of Arts and Sciences and of the American Statistical Association and a lifetime national associate of the National Academies. His work focuses on research methods for determining the severity and scope of social and education problems, implementation of programs and policies, and estimating the effects and the effectiveness of interventions. He contributes to work on randomized trials in education and training, welfare reform, health services, housing, and crime and justice, with a particular interest in the assessment or improvement of programs sponsored by federal agencies in the United States and by private foundations. He has a B.E. from Stevens Institute of Technology and a Ph.D. from Iowa State University.

Helen F. Ladd is the Edgar Thompson professor of public policy studies and professor of economics at Duke University. Most of her current research focuses on education policy. At the National Research Council, she cochaired the Committee on Education Finance, which issued two volumes: *Equity and Adequacy in Education Finance* and *Making Money Matter: Financing America's Schools*. She has written articles on charter schools, school-based accountability, market-based reforms in education, parental choice and competition, and a series of papers on teacher quality and student achievement. Currently she is continuing her research on teacher labor markets and teacher quality using North Carolina data as well as on various issues related to charter schools and parental choice of schools. Ladd has also written extensively on the fiscal implications of growth, property taxation, education finance, tax and expenditure limitations, intergovernmental aid, state economic development, and the fiscal problems of U.S. cities. She has been active in the National Tax Association (serving as president in 1993-1994) and the Association for Public Policy and Management and has consulted on tax policy and intergovernmental relations for all three levels of government. She has a B.A. from Wellesley College, an M.A. from the London School of Economics, and a Ph.D. in economics from Harvard University.

Martin Orland is director of evaluation and policy research at WestEd, a research and development organization specializing in education and human services. He leads a nationwide staff of methodologists, research scientists, content experts, and evaluators, with the responsibility for ensuring that their work employs the highest standards of methodological rigor and provides usable knowledge to decision makers. Previously he served as director of the Center for Education at the National Research Council. Orland has authored a number of publications for both academic journals and government and has regularly presented at international, national, and regional conferences in the fields of education and human development. He currently serves on the editorial board for *Education Finance and Policy* and has had two stints on the board of directors of the American Education Finance Association. He has B.A. and M.A. degrees, both in political science, from Brooklyn College, City University of New York, and a Ph.D. in social science from Syracuse University.

Jay Pfeiffer is deputy commissioner in the Division of Accountability, Research, and Measurement of the Florida Department of Education. He has been employed in the Florida government since 1972, beginning with the State Manpower Council. In 1984, he started the Florida Education and Training Placement Information Program, a first-of-its-kind program linking data from multiple administrative data resources for accountability purposes. In the Department of Education, he was the director of education information and accountability in the Division of Accountability, Research, and Management and, in February 2005, was promoted to his current position. He has a B.S. in chemistry/biology from the University of Florida, with graduate work in bacteriology and anthropology.

Barbara Schneider is the John A. Hannah distinguished university professor in the College of Education and the Department of Sociology at Michigan State University. She worked for 18 years at the University of Chicago, holding positions as professor in sociology and human development and as a senior researcher at the National Opinion Research Center (NORC). She continues to hold an appointment as a university faculty research associate at the University of Chicago and as senior fellow at NORC, where she is the principal investigator of the Data Research and Development Center. She also continues to direct the Alfred P. Sloan Center on Parents, Children, and Work at Michigan State University, an initiative that began at the University of Chicago, where she was codirector. She uses a sociological lens to understand societal conditions and interpersonal interactions that create norms and values that enhance human and social capital. Her research focuses on how the social contexts of schools and families influence the academic and social well-being of adolescents as they move into adulthood. She has published 12 books and numerous articles and reports on family, the social context of schooling, and the sociology of knowledge. She is the current editor of *Sociology of Education*. She has B.S. and M.A. degrees from National Louis University and a Ph.D. from Northwestern University.