Harnessing the Potential for Research of Existing Student Records Databases: An Action Agenda

National Center for Higher Education Management Systems (NCHEMS)

The Problem. The past decade has seen considerable progress in the development of state and system-level educational databases that contain information about student progress and academic achievement. This information is of interest to both the academic and policy research communities. Most of the relevant databases are administrative records systems maintained by public governing or coordinating authorities. Commonly termed "student unit record" (SUR) systems because they contain an individual electronic record corresponding to each student enrolled in an institution or school each term or year, they were originally developed to support enrollment-driven funding formulae and for official statistical reporting. State and system SUR systems are increasingly being mined by policy researchers in the states that maintain them to monitor enrollment flows, support policy studies, and construct accountability reports. But they are also of growing interest to the academic research community because their contents can support more sophisticated studies of broader social and educational issues.

Particular research questions that can be fruitfully examined using data drawn from such databases include (but are not limited to) the following:

- What does the overall "flow" of students through the nation's "educational pipeline" look like? Who succeeds in completing what kinds of credentials or outcomes in what periods of time, and how are these outcomes related to characteristics like gender, race/ethnicity, income, or social background?

- What experiences (curricular or environmental) affect student success in making progress through the "educational pipeline?" Are such experiences especially effective or inhibiting in particular settings or for particular student populations?

- What facilitates successful student transitions across specific experiential boundaries—for example from high school to college, from two-year colleges to four-year colleges, or from either of these to or from the workplace? How are these transitions different for different types of students? What programs, policies, or interventions seem to work best in helping students make such transitions and are their effects most pronounced in particular student populations?

- What is the return on investment for individual students and their families who invest tuition dollars, or for states and the federal government who provide financial aid? Does it matter where aid is invested (e.g. in public vs. independent institutions)? How are particular student populations affected?

- What role does geographic mobility play in inhibiting or enhancing educational credentialing or attainment? What effects do mobility and attainment patterns have on state or regional labor force composition and job markets?

As interest in investigating such topics has grown, SUR systems are being supplemented by databases constructed by academic researchers themselves, who typically draw data from one or more SUR systems and add specially-collected data and/or derived variables.

Both types of databases contain a wealth of information valuable to researchers, but they are not currently accessed to their full potential. State SUR systems are often large and intricate, with documentation that is difficult to negotiate for users without programming experience. As a result, their use is usually confined to analysts within the particular state or system board that maintains them. Specially-constructed research databases, in turn, are generally used only by the relatively small research community that created them. Furthermore, they may cease being maintained, or go out of existence entirely, at the conclusion of a project or when grant support disappears.

Systematic efforts to promote greater access for educational researchers to both types of databases would enable their considerable potential to be better realized. But such efforts will require initial support to develop an appropriate technical and organizational infrastructure that will enable efficient, equitable, and professionally responsible access to the data. The case for pursuing this line of work is based on the following:

- <u>There is a growing policy imperative to address educational pipeline issues</u>. One part of this imperative consist of the "K-16" initiatives now being pursued by many states. These concentrate on aligning K-12 and college standards, assessments, and finance policies to ensure smooth student transitions from one sector to the next (including the growing role of community colleges). Another dimension is the urgent need for educated workers to fuel the new economy of the 21st century, which will more than ever require increases in the number of knowledge workers able to function in a global context. Information about the "supply chain" for "educational capital" is thus of growing importance to individual states and for the nation as a whole. A final dimension of this imperative—poignantly revealed by the aftermath of hurricane Katrina—are the escalating material inequities that threaten the social fabric of the nation. Inequality of educational attainment is an integral part of this challenge. These are issues that are of supreme importance and that are quickly converging into an integrated policy agenda. Like the post-Sputnik agenda pursued by a mobilized scientific research community in the 1950s, this policy agenda could be fruitfully pursued by the social scientific and educational research communities in the coming decade if given the right tools.

- <u>Most of the important public policy and research questions related to this imperative can *only* be addressed through the use of these records</u>. Although national longitudinal datasets compiled by the federal government are available,

they are limited in scope and content, and they are only undertaken rarely. Because of limited sample sizes, such datasets cannot address state or regional questions, which are often the ones of most policy salience, because of limited sample sizes. At least as important, federal datasets cannot be linked to other data of interest—for example, data on workforce participation or student financial assistance—unless questions about these matters were asked as part of the original study. In contrast, current SUR databases contain millions of individually-identified student records covering the majority of educational activity in the nation. These records can be merged to create longitudinal files to investigate student success and behavior over time across settings and treatments. They can be disaggregated flexibly to examine impacts on particular populations. And they can be linked to other databases containing variables of interest via individual identifiers. Policy questions like the impact of particular state financial aid strategies on the success rates of low income students, the effectiveness of remediation or early-college programs in furthering K-16 objectives, or the impact of particular kinds of educational programs on local labor markets cannot be pursued in any other way.

- SUR databases have grown to the point that the educational records they contain constitute a "national asset" for research. Although these databases were originally conceived and built to serve quite limited purposes, their collective contents are now of enormous potential value beyond those purposes. This information now constitutes a public good in which many parties have an interest. The "open source" movement that has burgeoned over the last decade has made software and administrative systems widely available. It is grounded in the conviction that such assets are beneficial to many stakeholders and that these collective benefits should transcend (within reason) narrowly-construed proprietary interests. A similar philosophy seems warranted for the growing store of information contained in SUR databases. Although rightly "owned" by individual authorities, reasonable access to serve public purposes ought to be facilitated.

Taken together, these three conditions suggest that a deliberate effort be undertaken with joint foundation support to increase research access to current SUR databases and to begin to mobilize the academic research community to pursue topics of policy importance related to student progression. To explore what might be done to create such an initiative, a meeting of academic researchers and individuals responsible for several state SURs was convened on July 25-26 2005 by the National Center for Higher Education Management Systems (NCHEMS) with support from the Ford Foundation, the Lumina Foundation for Education, and the Spencer Foundation (see Appendix). The action agenda presented here resulted from that meeting.

Background. According to a recent inventory, a total of 46 SUR databases are in place in higher education in 39 states (Ewell, Schild, and Paulson 2003). Most only contain records of students enrolled in public institutions. But at least twelve now include data on students enrolled in independent colleges, and more states are planning to go this

direction. Collectively, such systems contain basic information about some 73% of the nation's headcount enrollment in colleges and universities. All have been in place long enough to track students longitudinally to the point of earning a degree (six years) and many have much longer histories. In addition, all have sufficiently common data contents and structures to support research on such prominent educational pipeline issues as retention and completion rates by gender and race/ethnicity, patterns of attendance at multiple institutions, or progression from remedial to collegiate study. Most do not, however, contain sufficient transcript-level detail to enable studies of course-taking or how curricular requirements are fulfilled.

According to another recent survey, 35 states have compiled sufficient student information on public K-12 enrollees in SUR databases to ground useful analytical work (http://www.nc4ea.org/index.cfm?pg=surveyresults). Many of these investments in information systems in the K-12 arena have been stimulated by the "No Child Left Behind" Act, which mandates reporting on student progress on a school-by-school basis and is providing financial support for states to improve their data systems. All of these databases contain information on basic enrollment, most contain data on tested proficiencies, and a growing number can be linked periodically to corresponding state SUR databases in higher education. Many of these K-12 SUR databases are already regularly tapped by organizations like Just for the Kids to create publicly-reported lists of high performing schools. The National Center for Educational Accountability (NCEA) has used them in a similar fashion to assemble a comprehensive five-state research database containing more than 12 million records.

About half the states have at least some experience with linking SUR databases for particular policy studies—for example linking postsecondary and K-12 data to look at preparation issues and the transition from high school to college, or linking higher education records to state Unemployment Insurance (UI) wage records to examine job placement and mobility for postsecondary graduates. To illustrate, the State Board of Community and Technical Colleges (SBCTC) in Washington is engaged in an ongoing program of research on the success of such programs as worker re-training, workforce development and adult basic education through linked SUR databases. In most cases, however, such linkages are accomplished to support special studies and the resulting merged datasets are not maintained once a given study is completed. A few states, however, have developed a comprehensive student tracking capability that encompasses the entire educational/workforce pipeline. Florida's Office of K-20 Education Information and Accountability, for example, is a state-level entity charged specifically with maintaining SUR records drawn from K-12, higher education, and wage records (as well as many others). Texas has created an equivalent capability to regularly link records from all three sources on a more decentralized basis, and several other states are quite close to developing this capacity.

Finally, a number of additional unit record databases that might be useful for educational and social researchers have been developed by various entities for specific purposes. For example, some university or association-based research teams have developed their own "value added" databases to support specific research programs. Databases like these

contain basic records extracted from state or system SUR databases, but are deliberately configured to support research by integrating data elements from multiple sources, creating derived or calculated data elements designed to answer particular research questions, and providing improved documentation so researchers can navigate them efficiently and effectively. One example is the dataset constructed by Marta Tienda and colleagues at Princeton for the Texas Higher Education Opportunity Project. Another is a longitudinal database constructed for a project based at USC entitled Transfer and Retention of Urban Community College Students (TRUCCS) which is based on many years of transcript-level records for students enrolled in the Los Angeles Community Colleges. Additional research databases like these have been constructed with foundation dollars to underwrite specific lines of research.

In a different vein, the National Student Clearinghouse (NSC) in Herndon Virginia maintains a limited set of data elements provided to the Clearinghouse directly by high schools and colleges. These are used to generate reports for clients on a fee-for-service basis that can be used to verify claims that a particular student has enrolled or graduated from a particular institution or to address such questions as whether students that have left a particular college or university have enrolled elsewhere. Growing numbers of states and institutions are using this database to generate accountability reports, but it has not been widely used to support scholarly research.

Issues Associated with Broadening Access to SURs for Researchers. Despite the considerable potential for mining these existing databases for broader research purposes, doing so requires attention to a number of practical and political issues. They include the following:

- Database Organization. Some state and system SUR databases are organized as relational databases or warehouses—an environment that allows individual data elements to be easily manipulated. Far more are organized simply as collections of flat (ASCII) files that are accessed and matched as needed, or are housed in a proprietary transactional software system that can only be manipulated by specialists. Furthermore, many SURs claim to be "warehouses" although they are neither fully relational nor easy for third parties to access and manipulate. And whatever the underlying data structure, few extant SURs contain the kinds of derived or analytical variables needed for research purposes.

- Definition and Documentation. Most state SUR databases in higher education use federal IPEDS definitions for a number of variables, but have established unique state definitions for many other things. State databases in K-12 education are somewhat less consistent, but are also achieving a measure of definitional commonality due to the impact of No Child Left Behind. More importantly, state SUR systems are typically documented for programmers, not analysts. In some cases this may simply pose a user challenge. But in others, lack of clarity about definitions might lead to serious misinterpretations. Finally, because state and system SUR databases depend on institutions to upload data from their own transactional records, the actual data contained in a given data element may not be

completely in accord with established definitions, or it may be missing entirely. More subtly, but at least as important, local conventions or "data cultures" may mean that different institutions handle similar data elements somewhat differently—a fact that will be known to analysts accustomed to working with these data within the state or system but that may mislead third parties who are ignorant of such conventions.

- Rules Governing Access. Third party use of any public database must be governed by agreed-upon rules that define who can get access to what data for what purposes. Some state SUR systems have multiple levels of users, each of which is allowed a different level access. Some states or systems also allow third parties to use such databases only for a particular period of time, after which any analytical files are destroyed. Still others require researchers to be physically present at a state or system office, working under the direct observation of database managers. In addition, access rules typically specify the "boundary conditions" of use—that is, what kinds of uses or applications are prohibited. Finally, though an access protocol may be clear and concise on paper, navigating it in practice may take researchers a considerable amount of time as various bureaucratic steps are cleared.

- Privacy/Confidentiality. "Public use" files derived from administrative records typically have all personal identifiers removed from individual records to protect privacy. But stripping individual identifiers means that no additional information can be subsequently linked to those records unless the original owner of the database accomplishes the merge. As a result, standard procedures specifying exactly who does what must be developed to perform tasks like updating longitudinal records. To maintain confidentiality, moreover, rules governing cell sizes are usually established so that individually-identifiable combinations of variables cannot be created and reported. There is also a good deal of uncertainty about how federal privacy laws such as FERPA bear on the third-party use or re-disclosure of student records to users not covered by any prior waiver of privacy rights. Finally, such entities as Institutional Review Boards (IRBs) are increasingly getting involved in adjudicating questions of data use beyond the original intent of a given data collection. Matters of privacy and confidentiality are complex for many reasons but one of the most important is that there are few consistent interpretations of what is "allowable" across jurisdictions. Different states have different interpretations of federal law and it is generally easier for state authorities simply to deny third party use entirely than to give this matter the detailed attention needed to comply.

- Organization and Governance. Successful third-party use of existing SUR databases requires an entity, or set of entities, to govern access. As third-party use has evolved for both policy and academic research, many examples of how such entities can be structured have emerged. In some cases, the entity in question actually houses the database. Examples include a single organization, a network of organizations, or consortia of users. But at minimum, such entities

must a) provide guidance on the contents of the various databases they have access to (together with examples of how these data have been used for research, as well as their strengths and limitations), b) provide information on how data can be accessed and the conditions of use and, c) provide some kind of feedback mechanism so users can give suggestions or communicate difficulties.

These issues not only pose challenges individually, but they also interact in complex ways. For example, database structure may decisively determine rules governing access independent of privacy concerns. On the other hand, an otherwise well documented and fully accessible database from a technical standpoint may be governed by access rules that are entirely driven by concerns about privacy or other considerations. These interactions—together with local political conditions—yield a distinctive context that surrounds each situation and that must be individually diagnosed and taken into account. Despite this caution, there are a number of actions that can be fruitfully pursued as part of a consistent national initiative.

An Action Agenda. A productive agenda for action to increase research access to SUR databases might be developed around six lines of work. Some of these are distinct and, to some extent, can be pursued independently through actions that can be undertaken immediately. Others will take more time and resources to develop.

1. Take Stock and Inventory. A first task in this agenda is simply to determine research demand, parallel efforts, and what data are available. Actions that might be undertaken consistent with this include:

    - Update and add to what is known about the contents of extant state SUR databases. With Lumina support, NCHEMS completed a fifty-state survey of such databases in 2002 (Ewell, Schild, and Paulson 2003). This survey will be updated in the fall of 2005 within the terms of an existing Lumina grant to NCHEMS. The 2002 survey focused on database content by examining each data element present with respect to definitional consistency and populations covered. The 2002 survey was also confined to postsecondary SUR databases, while the planned update will attempt to inventory K-12 SUR databases as well. In addition, the planned update will include information on a) reports regularly generated by the agency holding the database, b) experiences and capabilities with respect to linking data from other sources (and how such linkages were accomplished), c) state public records laws and additional state laws or regulations governing privacy or FERPA, d) procedures for gaining third-party access to data, e) previous experience with providing access to academic researchers and, f) ability to and procedures for contracting with third parties to conduct studies on behalf of the state and to receive funds for data services.

- Determine any parallel activities or plans of other potential players in this arena. Potential organizations here might include ETS, ACT, the College Board, active interstate compacts such as SREB or WICHE, and the National Student Clearinghouse.

- Identify additional scholars in the academic research community who are currently engaged (or might be interested) in work involving large student record databases. Most of the currently known researchers who have used SUR databases are in the fields of economics or higher education, but there are likely scholars in sociology, political science, or public policy who are also involved in such work. This process of identifying other interested scholars could first be pursed informally through personal contacts on the part of known researchers and funders, as well as contact with graduate training programs and research centers. Literature searches of the major social science journals might also be useful here. Once any initiatives consistent with this agenda are formally launched, moreover, it is likely that scholars will begin to self-identify in order to be part of it.

- Create a "Directory" Website. This would be a longer term initiative based on the findings of the previous activities. Its objective would be to create an initial "one-stop" website that could point researchers to available SUR data resources. Each available data source would be briefly described in a standard format with respect to data contents, access procedures, and previous research uses, together with direct links to those responsible for the database (and possibly other researchers who have used the database). A taxonomy of SURs should also be constructed that distinguishes a) sophisticated and well-developed state SURs where there is already a body of experience in serving researchers (e.g. TX, FL, OH, NC), b) "promising" state SURs which contain substantial bodies of information but that have not yet been documented or structured for third-party use (e.g. OK, KY, TN, HI), c) multi-campus university or community college system databases or large metropolitan school districts and, d) "value added" databases already created by researchers (e.g. Texas Higher Education Opportunity Project, National Center for Accountability, TRUCCS). Consideration might also be given to establishing a listserv or similar function to ease communication among researchers working on similar projects or accessing similar databases.

2. Launch and Document a Limited Number of Carefully Targeted Research Initiatives. This line of work should probably proceed in parallel with the inventorying activities described above. Its primary intent would be to discover lessons and create tools based on the experiences of selected scholars undertaking two or three actual projects in a limited range of database settings. A secondary goal would be demonstrate the utility of the process for other researchers. In addition to willingness to participate, preferred characteristics for a first set of demonstration initiatives should include the following:

- A "K-16" or "pipeline" perspective that involves the investigation of student transitions from one educational sector to another and/or the workplace.

- Inclusion of independent institutions in the study.

- Demographic coverage involving students drawn from multiple backgrounds and/or involving types of educational provision outside the mainstream (e.g. Adult Basic Education, ESL, GED, noncredit vocational instruction, etc.).

- Use of multivariate statistical research methods that are typically underutilized by state agency researchers.

Participating researchers would be work with and be accompanied by representatives of a third party organization with experience in large-scale student database design whose task would be to document the process to discover generalizable lessons and tools that might be applied to future projects. The third-party organization would also prepare a case study of each experience for broad dissemination that could also be posted on the proposed website.

3. Gradually Extend SUR Database Documentation. Using the lessons learned in these initial demonstration initiatives, a logical next step would be to begin to document additional SUR databases for inclusion on the proposed Website. This could be done on a case-by-case basis as research projects involving additional SURs are solicited and funded. It might also be done in parallel, without a research project in place, for databases that appear to provide promising settings for research and where preparing the needed documentation would not be a substantial task. Particular activities that might be undertaken under this heading include:

   - Create an "Annotated Data Element Dictionary" for each selected SUR. Like any Data Element Dictionary, these documents would include basic definitions and coding structures for all data elements in the selected database. But they should supplement these entries with annotations that describe a) cautions and embedded assumptions that might affect the use of these data for research purposes such as missing data or reliability issues, b) additional "local knowledge" about the limitations of these data such as special uses for particular codes or inconsistent applications of definitions by institutions and, c) the kinds of research questions that the element can be used to address with citations of previous research applications. These documents should also propose (or document) "derived" variables useful for research based on recodes or combinations of existing data elements. Each derived variable should be defined in terms of standard calculation procedures and the research applications for which they would be useful. A good model for the kind of documentation that should be provided is provided by the SOAR and ANSWERS websites that provide guidance about the large research databases compiled by NCES.

- Document any already-developed "data products" that may be available for each selected SUR. The objective here would be to take advantage of work already done by programmers and analysts associated with the SUR to create datasets that may be of immediate use to researchers. Such "data products" might include pre-built cohort files developed for regular longitudinal tracking and reporting, extract files with student identifiers removed that are already documented for public use, or "data cubes" designed to allow particular student outcomes to be disaggregated in complex ways without providing unit-record detail. Other "data products" might include analytical files developed to examine a particular problem or issues—for example, student performance in various skills areas, specific student populations, or student financial aid and tuition aid policies. If available and feasible, the actual code used to create and manipulate these datasets (e.g. SPSS, SAS, Access, etc,) should also be made available.

- Review relevant "best practices" associated SUR database design and implementation. The main intent of this activity, in contrast to the previous two, would be to assist in the development of SUR capacity in states that currently lack such capacity, or to help states with underdeveloped SUR systems improve them. "Best practices" in this case may include basic database design features (such as Web interfaces or relational structures), "value added" features that might facilitate researcher access, and tools for data manipulation and analysis.

The first two of these three activities would probably be most effective as enhancements to the proposed Website, while the third would likely be developed as a publication aimed at state data administrators.

4. Provide conceptual and "hands on" training for interested researchers in how to access and use targeted SUR databases. The basic intent here would be to develop a cadre of researchers in different disciplines who have the analytical and technical skills needed to conduct work using SURs. Two different kinds of activities appear appropriate here:

    - An Institute on longitudinal student data analysis held regularly at a major research university. The intent of this activity would be to provide basic conceptual training in how to approach various research tasks in the context of an existing SUR database. Topics that might be addressed include a) the basic architecture of typical files such as cohort files or enrollment-event files, b) the kinds of research questions that are typically pursued using SUR datasets and the particular research designs applicable to these questions, c) techniques for structuring and manipulating very large datasets including when and how to create derived variables or how to address multiple levels of analysis (e.g. student, class, institution, etc.), d) data analysis techniques with particular emphasis on multivariate techniques, and e) the "politics" of working in the

SUR environment including privacy issues and how to address them. The Institute would likely be at least a week in length and would be most effectively targeted at advanced graduate students or early-career faculty members. It should be held at a research university that has an existing cadre of researchers who are working in this area. A good model might be the two-week ICPR survey training workshops held each summer by the Survey Research Center at the University of Michigan in Ann Arbor.

- Focused workshops on the use of particular databases. These training opportunities would be intended to be shorter, more focused, and more "hands on" than the Institute and would be associated with a particular SUR. They would be held on an as-needed basis and hosted by those associated with the SUR database itself or by a research team that has created or worked with the SUR extensively. A three or four day format would probably be most appropriate, with topics including a) basic data contents of the SUR (using the previously compiled Annotated Data Element Dictionary) with particular reference to any embedded assumptions and limitations of specific data elements, b) access protocols and procedures associated with the targeted SUR, c) basic data manipulation techniques associated with this database and, d) reviews of past research projects that have been conducted using this database. At least half the workshop time should be devoted to exercises that involve direct manipulation of data in the specific database environment. Participants would probably benefit most from such training activities if they brought a particular research project idea to the workshop for practice and validation, and a proposal of this kind might be included as part of an application. A good model for such a workshop is provided by the summer workshop held in July 2005 by the Texas Higher Education Opportunity Project.

The first of these activities could be undertaken fairly quickly because the conceptual issues it addresses are for the most part matters of settled knowledge. The second would depend on the level of documentation available for each selected SUR, so would likely be designed later in the initiative. Participation in either or both kinds of training activities might be made a strong suggestion (or even a condition of funding) by foundations underwriting such work.

5. Develop an RFP/MOU Process Intended to Regularlize Relationships Between Database "Owners" and Research Users. Using experience gained from a growing set of demonstration projects (as described in #2 above), this activity would be targeted at developing a set of protocols and tools that could be quickly adopted by any state or district. These protocols would be based on the principle of "quid pro quo," involving careful specification of the benefits to and limitations placed on each party. The model RFP/MOU process should contain the following features:

- The state (or other database "owner") should specific a set of concrete deliverables in which it is interested for distribution to the research community. This might be done directly at a designated location on the proposed Website regularly updated by the Website administrator. Each posting would be in the form of a Request for Proposals (RFP) indicating state interest in a particular research question or deliverable, together with its priority from the perspective of the state. This would differ from a conventional RFP in that the state (or database "owner") would not be expected to underwrite the research, but would instead agree to provide access and orientation to the data. Researchers would be expected to provide funding or to seek it from extramural sources. Proposed deliverables might include a) specific research studies that address an area of need or interest to the state (e.g. how to successfully transition low-income students from K-12 settings to vocational/professional postsecondary programs with high income potential and high potential for in-state retention after training is completed), b) spin-off databases or datasets that the state can use for additional analysis and reporting but which are not within the current capacities of their own staffs to create, c) analytical tools and approaches (such as multivariate statistical tools) that researchers will develop in the course of their work but that state agency staff can apply to other projects after training by researchers and, d) joint work with agency staff that could result in publications or professional staff development.

- A formal Memorandum of Understanding (MOU) governing the relationship between the database "owner" and the researcher. The MOU should ensure that all work done by researchers be structured essentially as a contractual relationship in which the researcher acts as an "agent" of the state. Structuring the relationship in this manner will aid compliance with relevant FERPA and privacy requirements because access to student records does not involve "re-disclosure" of protected information to a third party. The MOU should also spell out in detail the physical procedures used to access and use the data that the researcher agrees to such as, for example, the use of secure terminals in the agency's own offices (where applicable), the destruction of all identifiable records after the research is completed, and how (or whether) non-identifiable files can be created for later analysis.

- Establishing a "University Service Center" in each interested SUR state. Such a Center would be tasked with providing a set of services to researchers interested in accessing the student data resources of that state and would be intended to provide a single, readily identifiable point of contact. Advantages of housing such a Center at a university would include familiarity to the research community and the ability to offload much of the work of "hosting" researchers from the state agencies who administer SURs, which many agencies find a distraction. A good model for this might be the North Carolina Research Data Center currently housed at the Terry Sanford Institute's Center for Child and Family Policy at Duke University.

The development of these protocols would be intentionally envisioned as an ongoing and iterative process, informed by actual experience with creating a growing set of research partnerships.

Some Cautions. Throughout the process of pursuing this action agenda, a number of important cautions should be borne in mind. These may affect the pace of the process, the order in which particular activities can or should be undertaken, and the tactics used to leverage researcher and policymaker interest. They include at least the following:

- The impact of privacy issues cannot be overestimated, and this is a moving target. Excessive agency or policymaker concern about privacy has the potential to block *all* third-party access to the information contained in SURs and the political salience of this issue is currently high. For example, a recent letter from the USDOE's Family Policy Compliance Office (which oversees FERPA) to the Kentucky Council on Postsecondary Education raised questions about the Council's data access policy which was developed last year especially to govern the use of SUR data in an agreement with NCHEMS and the Ohio Board of Regents. While the concerns noted in this letter are manageable, no previous questions had ever been raised. This is also a particularly volatile time for privacy issues in connection with educational records data because of an ongoing controversy over a recent proposal by NCES to launch a national student unit record system for postsecondary education. This proposal is still in play and is strongly opposed by prominent legislators on privacy grounds. The resulting "halo" of controversy is clearly affecting all discussions on database access, including those in states like Florida where such access is already a matter of policy. How this issue will play out nationally is at this point unknown, but it will likely have an impact on the proposed initiative.

- It is important to recognize that the politics of research access is more important than any technical issues. While privacy is chief among these concerns, there are many other political issues that may arise in particular contexts including institutional-state agency relationships (especially for independent institutions) and political sensitivity to particular research topics (especially those associated with equity). At the same time, every context will require building a strong local knowledge base about political and policy culture (as well as "information culture") that may take time to understand. This means not only that the necessary time needs to be built into the development of needed tools and protocols, but also that appropriate differences in the resulting products must be anticipated.

- Changes in instructional delivery will increasingly influence the design of student records systems. More and more instruction—especially in the postsecondary universe—is being delivered in non-traditional formats that are independent of established academic calendars and term structures. State SURs currently

accommodate such instruction by finding ways to associate start and end dates with traditional academic terms. But this is increasingly problematic for mastery-based programs where different students may complete the same material at different rates. Again, the impact of this phenomenon on SUR database design and access is not yet known. But any changes that may occur will probably have an important impact on proposed tools and protocols.

These cautions can all be addressed. But they emphasize the importance of a flexible and open-ended agenda in pursuing this work.

References

Ewell, P.T.; Schild, P.D.; and Paulson, K. (2003). *Following the Mobile Student: Can We Develop the Capacity for a Comprehensive Database to Assess Student Progression?* Indianapolis, IN: The Lumina Foundation for Education.

Appendix

The action agenda proposed in this paper was developed through a meeting of researchers and SUR database administrators held in Boulder CO on July 25-26, 2005 with support from the Ford Foundation, the Spencer Foundation, and the Lumina Foundation for Education.

Foundation representatives present began the meeting by noting their particular interests in the topic. For the Spencer Foundation, the primary motivation is the considerable potential of K-12 and postsecondary SURs to develop longitudinal databases already demonstrated by several funded projects. What is needed is a strategy to coordinate these efforts and learn from one another. For the Ford Foundation, an attraction is the benefit of using existing data rather than undertaking expensive projects to collect new data. Another is the potential of connecting the policy research and academic research communities, and the kinds of synergies these connections might yield. For the Lumina Foundation, longitudinal work lies at the heart of the Foundation's core mission of increasing postsecondary educational attainment for underserved students. For the Gates Foundation, a concern is extending to higher education the Foundation's established line of work on the educational success of low-income and minority students in K-12 settings. For the Hewlett Foundation with its particular focus on technology, a narrower interest is in the open access to educational resources—a natural outgrowth of the "open source" movement to create public use software.

The next section of the meeting involved discussion of the kinds of research being conducted using data drawn from SUR databases. A key point made here is that research has exhausted the possibilities for policy work provided by existing national longitudinal databases. The major policy action in both K-12 and postsecondary education is at the state level, and that is where researchers would like to focus their attention. The need for cross-state analyses and comparative analyses to answer certain policy questions was also noted. The major need is for an approach to access that minimizes "up front" costs to researchers and that makes access as "user friendly" as possible. Most of the research experience reported emphasized the intricacies of state SURs and the difficulties of harnessing them without a considerable amount of initial investment of time and effort. The kinds of research reported addressed many specific topics, but a general theme involved transitions from one educational sector to another and into the workforce—which requires drawing data from multiple databases and raises issues about linkages. An important technical point driving the access question is that journals (at least in economics) are now requiring research datasets to be made available to other researchers.

Discussion then turned to the characteristics of state SUR systems and the challenges that database administrators face in providing access to researchers. SURs are increasingly challenged by budget constraints as state budgets decline. They are difficult to maintain and some have very old architectures which require a good deal of programming attention to manipulate data. This may mean the researchers face a "revolving door" of data keepers with whom they have to deal, even in a single state, leading to the question

of how to institutionalize data capacity in a rational way. Researchers also need to understand the priorities under which state agencies operate, and that they will devote access and attention to research projects that are most aligned with their priorities. All states are under increasing pressure to protect privacy through FERPA and researchers need to be aware of these restrictions. States do differ in how they handle these issues, however. Some, like Florida, "firewall" the data and conduct analyses under the auspices of a single agency. Others, like Texas, allow access under supervised conditions through a supervised terminal. Researchers need to understand these rules of the game that are likely to be different in every state.

At the conclusion of the first day of the meeting, participants collectively brainstormed ideas about what might be done to address these conditions. These ideas were organized into a potential framework for an action agenda that was presented on the morning of the second day and refined for the balance of the meeting. The result was the action agenda portion of this paper.

A list of meeting participants is provided as Attachment A and the agenda for the meeting is provided as Attachment B.

Appendix A

PARTICIPANTS
FORD/LUMINA/SPENCER MEETING – JULY 25-26, 2005

Jorge Balan
Program Officer
The Ford Foundation
320 E. 43rd Street
New York, NY 10017
Phone: (212) 573-4661
Fax: (212) 351-3650
Email: j.balan@fordfound.org

Pamela Burdman
Program Officer, Education
The William and Flora Hewlett Foundation
2121 Sand Hill Road
Menlo Park, CA 94025
Phone: (650) 234-4500, Ext. 5654
Fax: (650) 234-1954
Email: pburdman@hewlett.org

Jennifer Delaney
Policy Analyst
National Center for Public Policy and
Higher Education
152 North Third Street, Suite 705
San Jose, CA 95112
Phone: (408) 792-3149
Fax: (408) 271-2697
Email: jdelaney@highereducation.org

Stephen L. DesJardins
Associate Professor
School of Education
University of Michigan
610 E. University Avenue
Ann Arbor, MI 48109-1259
Phone: (734) 647-1984
Fax: (734) 764-2510
Email: sdesj@umich.edu

Peter Ewell
Vice President
NCHEMS
3035 Center Green Drive, Suite 150
Boulder, CO 80301
Phone: (303) 497-0371
Fax: (303) 497-0338
Email: peter@nchems.org

Peter Bloch Garcia
Associate Program Officer
Research & Evaluation, Education
Bill & Melinda Gates Foundation
PO Box 23350
1551 Eastlake Avenue
Seattle, WA 98102
Phone: (206) 709-3208
Fax: (206) 709-3280
Email: peter.blochgarcia@
Gatesfoundation.org

Paul Goren
Vice President
The Spencer Foundation
875 N. Michigan Avenue, Suite 3930
Chicago, IL 60611
Phone: (312) 274-6519
Fax: (312) 337-0282
Email: vicepres@spencer.org

Jan Greenberg
General Counsel
Texas Higher Education Coordinating Board
1200 East Anderson
P.O. Box 12788
Austin, TX 78711
Phone: (512) 427-6143
Fax: (512) 427-6127
Email: jan.greenberg@thecb.state.tx.us

Linda Serra Hagedorn
Professor and Chair
College of Education
Department of Educational Administration
and Policy
University of Florida
PO Box 117049
Gainesville, FL 32611-7049
Phone: (352) 392-2391, Ext. 263
Fax: (352) 392-0038
Email: hagedorn@coe.ufl.edu

Eric A. Hanushek
Hoover Senior Fellow
Hoover Institution
Stanford University
Mail Code 6010
HHMB – Room 151
Stanford, CA 94305-6010
Phone: (650) 736-0942
Fax: (650) 723-1687
Email: hanushek@hoover.stanford.edu

Patrick Kelly
Senior Associate
NCHEMS
3035 Center Green Drive, Suite 150
Boulder, CO 80301
Phone: (303) 497-0307
Fax: (303) 497-0338
Email: patrick@nchems.org

Bridget Terry Long
Associate Professor
Graduate School of Education
Harvard University
Gutman 465
13 Appian Way, Longfellow Hall
Cambridge, MA 02138
Phone: (617) 496-4355
Fax: (617) 496-3095
Email: bridget_long@gse.harvard.edu

Dewayne Matthews
Senior Research Director
Lumina Foundation for Education
PO Box 1806
30 S. Meridian Street, Suite 700
Indianapolis, IN 46206-1806
Phone: (317) 951-5300
Fax: (317) 951-5063
Email: dmatthews@luminafoundation.org

Brian McCall
Professor of Human Resources and
Industrial Relations
Director, Human Resources Research Institute
University of Minnesota
3-251 Carlson School of Management
321 19th Avenue South
Minneapolis, MN 55455
Phone: (612) 624-9521
Fax: (612) 626-7785
Email: bmccall@csom.umn.edu

Michael McPherson
President
The Spencer Foundation
875 N. Michigan Avenue, Suite 3930
Chicago, IL 60611
Phone: (312) 274-6519
Fax: (312) 337-0282

John Milam
Executive Director
HigherEd.org, Inc.
21 South Kent Street
Winchester, VA 22601
Phone: (540) 722-6060
Fax: (540) 722-4225
Email: jmilam@highered.org

Jay J. Pfeiffer
Director
Office of K-20 Education Information
and Accountability
Florida Department of Education
325 West Gaines Street, Suite 844
Tallahassee, FL 32399-0400
Phone: (850) 245-0429
Fax: (850) 245-9069
Email: jay.pfeiffer@fldoe.org

Pat Windham
Associate Vice Chancellor for Evaluation
Division of Community Colleges
and Workforce Education
Florida Department of Education
325 West Gaines Street
Tallahassee, FL 32399-0400
Phone: (850) 245-9482
Fax: (850) 245-9525
Email: pat.windham@fldoe.org

Jill Wohlford
Program Analyst
Lumina Foundation for Education
PO Box 1806
Indianapolis, IN 46206-1806
Phone: (317) 951-5300
Fax: (317) 951-5063
Email: jwohlfor@luminafoundation.org

David Wright
Senior Research Analyst
SHEEO
3035 Center Green Drive, Suite 100
Boulder, CO 80301
Phone: (303) 541-1609
Fax: (303) 541-1639
Email: dwright@sheeo.org

Appendix B

Harnessing the Potential for Research of Existing Student Unit Record Databases

Meeting Agenda

Overview

The purpose of the meeting is to explore steps that might be taken to render large transactional/policy Student Unit Record (SUR) educational databases more accessible to researchers interested in conducting studies of the educational pipeline. Invited participants include academic research scholars, representatives from three SUR systems/states, and representatives of four foundations. The meeting will be held in Boulder CO on July 25-26 and will occur in a noon-to-noon format.

The general plan of the meeting is in two phases. The first day will be devoted primarily to identifying research questions, potential issues, and initial action steps. The second will be devoted to further amplifying action steps and discussing the details of how these might be pursued. The meeting will be followed by a working paper outlining further steps that might be taken and who would take them.

Monday July 25

12:00-1:00 Opening Lunch

1:00-1:30 Introductions and Review Purposes of Meeting

1:30-3:00 Research Questions

This discussion will focus primarily on the researchers to determine the particular kinds of questions that they (and others like them) are currently pursuing, or would like to pursue, with this kind of data. The discussion should conclude with a) a preliminary listing of the primary lines of research and research questions that are likely to emerge in this area and b) a preliminary listing of primary obstacles that currently inhibit using SUR databases to address them.

3:00-3:15 Break

3:15-4:15 Database Contents and Capabilities

This discussion will be largely descriptive, providing a brief overview of the three SUR systems/states represented at the meeting. Representatives will each introduce participants to the entity and data resources with which they are associated. These brief

presentations will emphasize a) their assessment of the suitability of their databases to address the questions posed and, b) their previous experience in granting access to them for academic researchers. This session should also include some discussion of the implications of the NCES proposal for a federal SUR system to support IPEDS.

4:15-5:15 Issues Involved in SUR Database Access

This session will be a general discussion of the issues raised in the working paper prepared by NCHEMS (but will not be limited to these issues). NCHEMS staff will take running notes on potential action steps that might be taken to address these issues as they arise in the discussion.

5:15-5:30 Wrap-Up for Day

This brief session will summarize the potential action steps identified, providing a starting point for the next day's discussion.

Tuesday July 26

8:45-9:00 Revisit Potential Actions

The day will begin with a brief review of the potential action steps identified through the previous day's discussion.

9:00-10:30 Elaborate Potential Actions

This session will provide the opportunity to more thoroughly discuss each action step noted. The objective will be to flesh out each of these ideas in sufficient detail that it might constitute the basis for a fundable demonstration project or a model that could be applied to other SUR databases.

10:30-10:45 Break

10:45-11:30 Foundation Reactions

This session will provide an opportunity for the foundation representatives to reflect on what they have heard and provide feedback on what might be possible and fundable, short term and long term. If specific project ideas are sufficiently developed at this point, there might be some discussion of joint action to pursue them.

11:30-12:00 Wrap Up and Next Steps

This concluding session will summarize what was learned and to discuss who should do what as a result of the meeting. Specific attention should be devoted to how to broaden the conversation—particularly, which additional members of the academic research

community should be involved and what additional SUR databases might be useful to pursue.