The Power of Psychology's Databases

January 22, 2002

Why share psychological data? Aren't there serious obstacles that prevent sharing? And even if we were willing to share in principle, how would we go about doing it in fact? These questions were discussed in the first three articles in this series. The answers I have suggested are these: We should share for our own benefit, in order to advance our science, and to better serve the general public.

There are no obstacles to data sharing that have not either been overcome or that have no possible resolution. And despite psychological science's slowness to embrace data sharing, there exist several well-developed plans for how to do it on a larger scale than is currently the case. The final installment of this series looks at some of the psychology databases that already exist. Their variety, size, and accessibility are impressive. The databases discussed here are by no means the only psychology databases that are available.

Pioneers have been at work building what can become the most important new tools at our disposal for advancing understanding of human behavior. In the following pages, several databases are profiled: Henry A. Murray Research Center of Radcliffe, Archives of the History of Psychology, fMRI Data Center, National Archive of Computerized Data on Aging, and National Longitudinal Study of Adolescent Health.

A FINAL WORD

It remains to be seen whether we can answer those and other questions in ways that turn psychologists toward much greater acceptance of, and reliance on, shared data. There are at least two trends that may be distinguished. Each offers hope. On the one hand, we have the experience of the fMRI Data Center where the database wizards are struggling with how to present discrete, very large datasets in a single area in ways that will assure their utility. On the other hand, we have experiences such as the Murray Center and the ICPSR-managed archives where the struggle to achieve utility has an inward and an outward aspect. Not only are these centers working to make behavioral science data useful for a large potential user community (the inward aspect), they are also working at the problem of making those datasets compatible with data from many other disciplines (the outward aspect).Unlocking the power of behavioral science databases requires doing both things well. We need to get better at making individual experiments within single disciplines relate more comparably to each other. And we need to make the research of psychology mesh better with the related data of other sciences. The prize for solving these problems is a leap in the speed with which we accumulate scientific knowledge. There is reason to think the questions will be satisfactorily resolved.

Behavioral Science vs. Social Science Databases

Social scientists were building and using databases decades before psychologists even started to think about using them. In fact, scientists in some social science fields have long sacrificed funds for

individual research so that the funds could be pooled to support common data collection efforts. Understandably, the social science databases are more developed than are the behavioral science databases. Psychologists, however, are unlikely any time soon to sacrifice funds for individual experiments in favor of collective efforts.

There is another difference between the two kinds of databases that makes it hard to simply borrow the accumulated knowledge of social scientists in this area. Social science data are largely survey based and descriptive, and there is widespread agreement among users about the protocols under which the data are collected and made available. A substantial portion of psychological data comes from surveys, and it is worth noting that data of this sort have been the first to go into publicly accessible behavioral science databases. But the bulk of psychological data are experimental. Agreements about protocols from one experiment to another are, at best, loose. So there are questions that behavioral scientists face in putting together their databases that are directly related to the nature of the data to be archived and that differ from the questions social scientists need to ask to achieve orderliness in their archives.

Henry A. Murray Research Center of Radcliffe

The Murray Center (www.murray.harvard.edu/), founded in 1976, is home to over 270 datasets. Henry A. Murray, the personality theorist after whom the Center is named, devoted his career to the study of individual lives. The datasets he and those who have followed him at the Murray Center have collected reflect that bent. Ever wonder what happened to Lewis Terman's data once his classic Life Cycle Study of Children with High Ability was finished? They're alive and well in Cambridge, but lonely to meet new friends. They reside with the data from Mary Ainsworth's Baltimore Longitudinal Study of Attachment and Sears, Maccoby and Levin's Patterns of Child Rearing to name just a few of the all-time classics. They keep good company with more recent datasets such as Anne Petersen's Adolescent Mental Health Study, Kay Deaux and Joseph Ullman's Blue Collar Workers in the Steel Mill, Jacquelynne Eccles' Career Aspirations Among Smith Undergraduates: A Longitudinal Study, and Plomin, DeFries and Fulker's Colorado Adoption Project.

The data – not the journal articles, but the actual data – on which the careers of some of the most hailed scientists in our field were built are to be seen at the Murray Center. And these aren't dead data under glass we're talking about here. The data are there for your use, to be asked questions Eleanor Maccoby and Robert Plomin and Anne Petersen and Jacquelynne Eccles haven't thought of yet. More than that, it is possible with some of the datasets to gain access to the original respondents to ask new questions. Imagine that. Others took care of the first forty years of data collection for you so that you could simply build on their legacy.

It gets better. Henry Murray's philosophy was that a multidisciplinary approach is needed to understand the individual. While the common thread through all the Murray Center datasets is that they are tools to help us understand the individual, the data themselves are drawn from a variety of areas, not just from psychology but from psychiatry, anthropology, sociology, political science, indeed, the spectrum of the behavioral and social sciences. One might rightfully think of the Murray Center as an El Dorado: The gold is there for the taking.

How do you get to it? Easy. Go to the URL listed above. Choose "Data Archive" from the menu, and

then select “Accessing Data.” Click on “Registration Form” in the next window. Register. Click on “Application for the Use of Data” or “Computer Data Request Form” depending on the dataset or sets you want to use. Tell the folks at the Center what you’ve got in mind. They’ll mull it over for a few days or a few weeks depending on what you are asking to do. Once they approve, you’re in. And don’t worry. They will work with you to see that your experience at the Murray Center is a successful one. They are there to give psychology away.

Archives of the History of Psychology

Also steeped in the classics, yet a very different database from that at the Murray Center, is the Archives of the History of Psychology in Akron, Ohio. To see what’s available there, go to www.uakron.edu/ahap/. As its name implies, the Akron Archive is dedicated to promoting research in the history of psychology. If the Murray Center is psychology’s El Dorado, then the Akron Archive is our Smithsonian, that is to say, our attic.

Much of what is contained in Akron cannot be accessed on the Internet because it is solid stuff. Seven of B. F. Skinner’s teaching machines are housed there, for example. Is there an introductory psychology student in the last 40 years who has not seen the black and white film of Stanley Milgram’s startling study of compliance? The simulated shock generator that remains the iconic representation of that experiment is in Akron. There are other symbols of our history there too. David Wechsler’s psychogalvanograph and his Brunzwigia calculating machine are there. There is a film of Kurt Lewin explaining Field Theory (not to mention child development and the social climate of groups), and one of Ivan Pavlov concerning functions of the brain. What was Sigmund Freud like as a teacher? There is a reel-to-reel tape containing recollections of one of his students to give you some idea. What was Abraham Maslow like as a student? His class notes are preserved at the Akron Archive, as is Spence’s correspondence with Hull regarding the Hull-Spence Theory.

Since it opened in 1965, the Archive has been the repository for the records of many of the country’s psychological science societies. Our institutional memory literally resides there. In fact, when the Archive’s capacity to take in new material from the societies was exceeded several years ago and it was announced that no new material of that kind would be accepted, it sparked a serious discussion among leaders of the country’s societies about how the records of our organizations can continue to be preserved. It is a discussion that is still underway.

The Akron Archive’s travails over capacity and resources mirror that at many of our archives. With more resources, much more of the printed, filmed, and audiotape holdings of the Akron Archive could become available online for teaching and research. With more resources, people could be hired to put more of the papers of our most revered scientists in order and make them available for scholarship. For now, there is enough light in the attic only to show us the shadowed outline of our past.

fMRI Data Center

So far, we have a gold mine and an attic. Next we look at psychology’s brain, or at least a site where there are a lot of brain images that psychologists ought to be looking at. The functional Magnetic Resonance Imaging Data Center at Dartmouth College (www.fmridc.org) is one or our newer datasets.

Thanks to a substantial grant from the National Science Foundation, the fMRI Data Center opened its virtual doors in the autumn of 1999. A team led by psychologist Michael Gazzaniga, one of our foremost cognitive neuroscientists, operates the Center.

The Center presents a fascinating case study for those who would build a database with very large storage needs. In the heady, early days of thinking about the Center, it was believed that it should house the brain images from all the top neuroscience journals.

And maybe it shall. But not immediately.

Right now the core of the Center's collection consists of the data underlying the 13 fMRI studies published in the special supplement to the November 2000 issue of the Journal of Cognitive Neuroscience. Make no mistake; these data are the core of a collection that will grow steadily. But one of the things the database wizards have found is that to store neurological images, the space required is not to be measured in mere megabytes, nor even in the gigabytes which excite us commoners today, but in the multiple terabytes.

Scientists at the Center are in their early years of figuring out how to accomplish the physical task of storing and making such huge amounts of information available. Brain images are quickly becoming important to many areas of psychological investigation. Unfortunately, not every psychologist has a brain imaging machine at her or his disposal. And those who do have access often pay a pretty penny for the staff and machine time necessary to make their images. There are few dedicated research fMRI machines in the country. Instead, most of these machines are to be found in medical facilities and are used mainly for medical purposes. Scientists who wish to use the machines often find themselves doing so on the "off hours" and paying a premium for the privilege.

The need for images and the information underlying them outstrips the current capacity to produce them, not to mention the current expertise in imaging within the psychological research community. Enter the fMRI Data Center. Just as data from longitudinal studies may be used to answer questions the original investigators did not think to ask, so it is with brain images and the additional data underlying those images. It is important, therefore, that the Center's holdings include all the data that underlie the published fMRI studies that make up its corpus. That includes the pre-processed images and all the technical information that gives meaning to the images. Recognizing that many who need to use these data are not yet well versed in how to use them, the Center provides training in addition to providing data. Pretty smart. But that's what you'd expect from people whose business is the functioning of the human brain.

For now, you can use your Internet connection to look at the holdings of the Center. When you have settled on the data you wish to have, you submit your request and the Center ships you the data you have requested. No doubt the day will come when such large volumes of data can be sent easily over the Internet.

National Archive of Computerized Data on Aging

A few years ago officials at the National Institute on Aging asked themselves what the knowledge yield

had been from their decades of support for research on aging. While there were positive answers to the question, the officials also found that there were contradictions among the findings from NIA supported-research, and there were substantial gaps in knowledge. They wondered what they might do in the future to bring more consistency to the research results and to spot the gaps in knowledge more easily.

One of the solutions they settled on was data sharing, or more precisely, the creation of a data infrastructure for gerontological research. If researchers could have access to the data of fellow researchers, they would have a tool for resolving differences in findings that is not afforded by published articles alone. They would be able to exploit hard-to-collect but underutilized datasets. Moreover, greater availability of data within an ordered infrastructure could create a synergy among researchers that might accelerate the acquisition of knowledge about aging.

NIA entered into a partnership with the Inter-University Consortium for Political and Social Research (ICPSR), the organization that has long been the umbrella organization for database managers in the social (not the psychological) sciences. The result is the National Archive of Computerized Data on Aging (www.icpsr.umich.edu/NACDA).

NACDA performs three important services. It actively seeks out datasets and adds them to the archive. It processes the data, putting them into a form that makes them easy for researchers to use and that makes it possible to relate one dataset to another. And it provides the technical support scientists need to make effective use of this valuable and growing data resource. An important feature of the way NACDA operates is that it has a council of stakeholders from the sciences that contribute and use the data. The NACDA Council helps make decisions about such things as what to acquire and how to evolve the archive in ways that assure fulfillment of its mission to advance research on aging by helping researchers to profit from the under-exploited potential of a broad range of datasets.

The philosophy of NACDA is much like that at the Murray Center: To know aging, a multidisciplinary approach is needed. Thus, NACDA contains data from medical science, demography, economics, sociology, psychology, and many other disciplines as well. More than 100 of NACDA's datasets are free and publicly accessible. Because ICPSR is an umbrella organization for database managers, it is also possible to gain access to over 500 additional datasets not directly managed by ICPSR.

Policies governing the use of these datasets vary, and use of some of them requires paying a fee. Moreover, access to many of the datasets is restricted to ICPSR members. If your institution is an ICPSR affiliate, then you have the potential to use these other datasets. To find out if your institution is part of the club, go to www.icpsr.umich.edu/MEMBERSHIP/ors.html.

NACDA is also a gateway to datasets that are located neither in NACDA nor in an ICPSR-affiliated archive. Again, policies will differ with the vendor. But NACDA is your portal to vast stores of easy-to-use data on aging. ICPSR also manages the Substance Abuse and Mental Health Data Archive (www.icpsr.umich.edu/SAMHDA). Its policies and even the look and feel of the database site are similar to NACDA's. In fact, ICPSR's effort to give its databases a common front-end look is one of the user-friendly design considerations that perhaps should be more generally applied to the data archives of the behavioral and social sciences.

National Longitudinal Study of Adolescent Health

The database of the National Longitudinal Study of Adolescent Health (www.cpc.unc.edu/addhealth) was dearly won. The intention to create this database grew out of a realization on the part of researchers and policy makers that many public health decisions having to do with sexually transmitted diseases were being made on the basis of data collected half a century ago through the Kinsey sex surveys. It seemed absurd to be relying on data so old that it preexists AIDS to make decisions affecting the health and lives of people currently at risk for sexually transmitted diseases.

At the same time, these researchers and policy makers realized that there had never been a comprehensive survey of the health practices of adolescents.

These two needs drove a nearly decade-long fight in Congress to obtain funds to undertake the survey. Conservative lawmakers believed passionately that asking questions about the sexual lives of adolescents was way beyond anything government should sanction, let alone fund. So, year after year, money was placed in appropriation legislation to fund the study. And, year after year, the money was removed.

Many modifications in the research design were made, and, eventually, the go ahead was given to conduct the study. That some of the outcomes of the study seem to support conservative positions regarding sexuality had a noticeably positive effect on conservative support for the research. The survey, however, demonstrates in the best tradition of science that the outcomes of research are not set in advance.

To date, two waves of data collection have taken place. Wave I data were collected between September 1994 and December 1995. Wave II data were collected between April and August 1996. The survey is a school-based study of the health-related behaviors of adolescents in grades 7-12. It assesses the health status of participants and also explores the causes and contexts of their health-related behaviors.

Three contexts are examined: the social context, the personal decisions and practices themselves, and the personal and personality variables associated with particular behavioral decisions. Those three levels of data make this the richest dataset available for understanding the health behaviors of adolescents.

The sample size is over 90,000 adolescents from 80 communities chosen so as to make the sample representative of the U.S. by region, school type, urbanicity, ethnicity and school size. All respondents completed an in-school questionnaire. From these respondents, a core sample of 12,105 adolescents was taken. For these individuals, an additional in-home survey was conducted. Additional special oversamples of certain ethnic and racial groups were taken as well. There is also a large sample of adolescents with physical handicaps.

The questions it is possible to ask of these data are nearly limitless, and your opportunity to be one of those asking the questions is very real. The data are available in two forms-a public-use dataset and a restricted-access, contractual dataset. The sensitivity of these data makes protection of the respondents of cardinal importance. So the public use dataset contains only a subset of respondents. To gain access to the restricted-access data, you must be a certified researcher, and you must commit yourself to

maintaining limited access. Deductive disclosure, that is, the ascertaining of the identity of a respondent by bringing together disparate pieces of data, can occur with this dataset. That is why anyone who uses the restricted dataset is obligated to protect respondents from that eventuality by submitting to a number of precautionary actions.

Don’t be intimidated by the importance of the data. The Add Health Internet site is very well organized for the researcher, and the principal investigators are eager for researchers to wring the last scintilla of knowledge from the data. There is much more to be learned from these data than the causes of the health behaviors of adolescents. What do you want to know? The Add Health database may have your answer.