Three Objections to Databases Answered

November 27, 2001

Sixteen years ago, the National Academy of Sciences published a report on data archiving for the behavioral sciences entitled *Sharing Research Data*. Much of the report was devoted to a discussion of objections by behavioral scientists to having databases. In the years since the report was published, the effort to create data archives for psychology has limped along without generating widespread enthusiasm, and the objections to databases have remained unchanged. That lack of change is remarkable because experiences with the databases that have managed to come into being have demonstrated that workable answers exist to each of these objections.

A database is any repository for information. The term is often used interchangeably with the term, "archive." In this series of articles, these terms refer to electronically stored and accessed repositories. Ideally, the information contained in these repositories is assumed to be of three kinds: raw data from psychological research; information regarding the nature of the data (which we have termed "metadata"; and information about unique aspects of the database (which we have termed the "contingency table."

What follows is a discussion of the three main classes of problems that have been noted, with an emphasis on answering those objections. The problems are described in terms of the foci of the problem areas, which are research participants, science, and the careers of scientists.

The Participant Problem:
Databases will compromise the right to confidentiality of those who participate in psychological research.

RESPONSE: This is a serious issue; preserving confidentiality is a critical element in psychology research. But there doesn't have to be a tradeoff between confidentiality and access. Several approaches to maintaining confidentiality have proven successful. The Adolescent Health Survey (for further information, go to www.cpc.unc.edu/addhealth) is a good example of confidentiality protection through physical *separation of identifiers* from data. The survey was conducted several years ago after half a decade of heated debate in Congress about confidentiality and the highly personal questions about sexual practices and other aspects of adolescent health contained in the survey. Although the data are publicly available for further analyses, all information that could be used to associate individuals with their responses has been removed and stored separately from the data.

A second approach has been to introduce *controlled error* into the data. In analyses, the error can be compensated for mathematically such that results are not hopelessly distorted. But, in effect, the data that are available for analyses are merely based on the responses of participants. The dataset may consist of all the responses of each participant, but it is not composed of only those responses. The dataset is real in the sense that responses were collected from real people, but it is also virtual in the sense that no

response set is a wholly accurate representation of the responses of any participant in the study.

This ingenious approach to confidentiality was developed by Willenborg, Fienberg and de Waal of the Department of Statistical Methods at Statistics Netherlands (Willenborg & deWaal, 2001; Fienberg & Willenborg, 1998).[1]

Aggregation of data is a third approach to guaranteeing confidentiality. This is a method that has been used by the National Center for Education Statistics (NCES) since it began making its datasets available on CD-ROM. As the name implies, this method combines individual data into group data. There are levels of aggregation. Under the NCES system, access to data is restricted by degrees. That is, at the most aggregated level, the data are available to anyone for the asking. Access to less and less aggregated is governed by greater and greater restrictions on the user. At the least aggregated level, the user must justify the need for the data, must sign a legal document agreeing to the terms of use, and must come in person to the offices of NCES in order to use the data under NCES personnel supervision.

Not all challenges to confidentiality have been solved. But there are a number of potential solutions. For example, current informed consent procedures may contribute to the problem of providing access to data while preserving confidentiality. Some have argued that the wording of informed consent agreements precludes the release of data to anyone except the principal investigators even if responses cannot be traced back to individuals. Addressing this problem may warrant a regulatory change; it may become necessary to amend the relevant regulations in order to separate some aspects of access to the data from guarantees of confidentiality within informed consent procedures. Future consent agreements could guarantee the participant that his or her responses will never be traceable, but would also make clear that others besides the principal investigators may have access to data from the study. The most likely target for such a change would be a document called the “Common Rule” (Code of Federal Regulations, Title 45, Part 46) which spells out the requirements that federal agencies that are subscribers to the Rule must demand of grantees who use human participants in their research.

The second unsolved threat to confidentiality arises in connection with data that become useless unless some aspects of the identities of the participants are known. Videotaped material is a case in point. Developmental psychologists have used such tapes in their efforts to understand childhood behaviors. Facial expressions are essential indicators of behavior, and the proper coding of observed behaviors requires being able to see the child’s face clearly.

In the short run, removing place and time identifiers from the data can make it difficult to discover who the participants are even if their faces can be seen. In the longer run, technological fixes may become available. For example, producers of animated movies now use a process that begins with the movements and expressions of real people. Those movements and expressions are transferred faithfully to characters created by computers. The process is costly now. But it may one day be possible to transfer to avatars all the scientifically interesting content of behavior captured on videotape.

While the concern about confidentiality is probably the most often expressed reservation about publicly available databases in psychology, there are few such concerns that cannot and indeed, have not been addressed satisfactorily.

The Science Problem:

It is wrong to treat discontinuous data as though they were continuous.

RESPONSE: When meta-analysts first began lumping together bodies of research and treating results like one huge study, there was great objection in the community. People said, and some still say, that this is comparing outcomes that cannot validly be compared because the data – even those from related studies – are collected at different times, under different circumstances, by different people with different views and different purposes.

It should go without saying that, despite the objection, this is exactly what we do every day. And we do it without the exactitude that could be afforded by access to raw data rather than just reported results.

How many disputes in psychology are, in actuality, misunderstandings of nuances of methodology? How often are results artifacts of paradigms rather than true indicators of the causal bases of behaviors? Currently, we debate the answers to such questions in something of a vacuum. With only reported results on which to base the arguments, we are handicapping the general effort in psychological science to build a reliable body of knowledge.

Databases, in defiance of the criticism against them, are proving to be important test beds for methods to improve the comparability of one study to another. Here are two cases in point:

For the past twenty years, John J. McArdle[2] of the University of Virginia, has been collecting datasets from administration of aptitude and intelligence tests. Beginning with the Wechsler Adult Intelligence Scales, he has recently added the Woodcock-Johnson scales to his database.

McArdle has amassed a database of literally hundreds of thousands of test results spanning many decades of testing. Among the items in the database are norming data from some of the most used tests of ability, and testing results on cohorts of participants in a variety of longitudinal studies. These are data that were formerly inaccessible to any except the original principal investigators or personnel of test development companies. McArdle has been using this rich archive to create revolutionary new ways to analyze datasets that were formerly thought to be so different as to defy comparison. Not only is the database he continues to assemble of great value for the raw data it contains, but the use of the database to test new approaches to mathematical comparison is giving important new analytical tools to psychological scientists.

Similarly, developmental psychologists have been putting together an archive of recorded child behaviors. The database began as a collection of aural recordings of children in a variety of interactive settings. More recently the sound recordings have been supplemented with videotaped material.

Originally, the developmentalists began to share these data because it was economical to share a precious resource; they realized that limited financial resources meant that some hard-to-obtain data could not be collected *de novo* each time a new researcher needed such data. But as time went on, the database users found that they could use the recordings as a means to develop standardized behavioral descriptors. It had been a problem in observational studies of child behavior that different researchers coded the same behavior differently from one study to another. Using the taped material, developmental

psychologists have now developed standardized coding procedures that have brought unprecedented coherence to research on child behaviors.

Far from leading the field into false comparisons of data, the experience we have from current work on databases suggests that in the right hands, databases are becoming means to achieve more rather than less reliable comparisons, and greater rather than less agreement about the validity of scientific knowledge from psychological research.

The Scientific Career Problem:
Those who contribute their work to databases supply their critics with the means to damage their careers.

RESPONSE: This problem has been voiced in several forms: 'I worked hard for my data. Why should I give it away?' 'No one who wasn't in on the designing of my study and the decisions regarding analyses of the data will understand the data well enough to use it properly.' 'My critics will just misuse my data for their own ends, hurting me in the process.'[3]

In the scientific community, psychologists are notorious for their skepticism about each other's work. So the scientific career problem with respect to databases goes to the heart of the culture in which psychological research is carried out, and is thus to be taken most seriously. But, our culture notwithstanding, the problem of misuse of archived data is not the sole province of psychologists. It is a general problem that has been addressed on an objective level and solved by those in sciences that have long made use of databases.

Those who have addressed the problem have noted that a database is not simply a collection of the data points themselves. If a database is to be useful for secondary analyses or for answering new scientific questions it should have a tripartite structure. It should be composed of the raw data, a meta-database, and contingency descriptors. Little needs to be said of the raw data. They are the numbers or survey responses or other measurement units that are proxies for the behavior under study. The meta-database, however, is less familiar to psychologists, although it should not be. It is the amplified equivalent of the methods section of a journal article. It is the detailed description of the methodology, assumptions, coding procedures, analytical decisions, etc. that formed the general parameters within which the data were collected.

The contingency descriptors are the special notes that link the general method with the particular data obtained. They describe the unique aspects of the study such as why some data points were removed from the database, or why unexpected changes in testing occurred. A database structured in this fashion not only allows secondary analyses and in-depth assessments of the data; it also can be a scientist's best guarantee that his or her work will not be misinterpreted. Currently, disputes about outcomes revolve around the report of results as published in journals, not around the data themselves. The tripartite database, as described here, would provide the common groun on which researcher and critic could disagree based on substance rather than misunderstanding. Why? Because all parties would have access to the raw data, the detailed methodology, and the description of occurrences that may have played a role in producing the data. The general availability of the data, the meta-data and the contingencies

makes it much easier for the larger research community to judge the place of the research in the overall effort of the discipline to increase understanding of human behavioral determinants.

So, databases have the potential to do the opposite of what critics contend they will do. They can preserve confidentiality and, at the same time, maximize access to data. They can bring greater integration of scientific knowledge rather than chaos borne of false comparisons, and they can be the guarantors, rather than the means to ruination, of scientific reputations and careers. Databases are crucial to the rapid advancement of psychological science. In the end, however, it may be fear of ourselves that presents the greatest impediment to the building of databases and that most undermines the quest to integrate knowledge in psychological research. Our wish to avoid embarrassment by continuing to present only our interpretation of results rather than the data points that underlie those results could have the effect of depriving the science of a tool it sorely needs.

Next time: Three approaches to database building.

REFERENCES

Fienberg, SE, Martin, ME, Straf, ML, eds. (1985) Sharing Research Data. National Academy Press, Washington, D.C.
Fienberg, SE & Willenborg, L. (1998) Special issue on "Disclosure limitation methods for protecting confidentiality of statistical data." Journal of Official Statistics, 14 (4), 337-566.
Willenborg, L. & de Waal, T. (2001) Elements of statistical disclosure control. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

---

[1]Much of their work is available, at least in summary, at this Statistics Netherlands website: www.cbs.nl/sdc/.

[2]McArdle is a quantitative psychologist with an impressive list of honors that have included the presidencies of the Society for Multivariate Experimental Psychology and the Federation of Behavioral, Psychological and Cognitive Sciences. He is currently Secretary of the Council of Scientific Society Presidents, chairs the Evaluation Research Committee of NIH's Center for Scientific Review, and serves on the Institutional Review Board Panel of the National Research Council.

[3]One prominent researcher has been quoted as saying that the analysis of data is a right that should be earned through participation in the competitive granting process, and that those who carry out secondary analyses of the data of others without their permission or oversight are "carpetbaggers."