Journal of Intercollegiate Sport, 2009, 1, 5-8

A Process for Sharing Research Data Collected by the NCAA

Todd A. Petr and Thomas S. Paskus

National Collegiate Athletic Association

The importance of managing, archiving and potentially sharing large-scale data collections has become salient in many academic fields of study during the past decade. Technological innovations have quickly enhanced our ability to manage and archive data. However, the sharing of data among social sciences researchers has remained a rather limited practice (Breckler, 2009; Freese, 2007a), despite some evidence showing that benefits accrue even to the researchers providing the data to others (Gleditsch, Metelits, & Strand, 2003). Although some funding agencies require (e.g., the National Institutes of Health) or encourage (e.g., the National Science Foundation) data sharing in certain circumstances and most social science disciplines express broad support for the practice in one form or another (Freese, 2007b; King, 2007), movement toward routinely sharing data has been glacial in education and psychology among others fields (see Azar, 1999; Breckler, 2009; DeAngelis, 2004). Certainly, concerns about the confidentiality of research participants, the time and cost involved in preparing data for broad dissemination and the desire to fully mine data that may represent a substantial financial and intellectual investment all play into data-sharing hurdles faced in these areas of study. At the same time, most would agree that a discipline benefits when substantial or unique data are made available to other qualified researchers.

The study of intercollegiate sport is a discipline that would likely benefit substantially from an enhanced commitment to sharing research data. This was noted during the first annual Scholarly Colloquium at the National Collegiate Athletic Association (NCAA) Convention in January, 2008. Jay Coakley presented a broad review of factors affecting research in the area of intercollegiate athletics, and the ability of scholars to conduct such research effectively (Coakley, 2008). One of the issues that Coakley raised was the wealth of quantitative data collected by the NCAA to answer research questions posed by its members and assist in the development of national athletics policies. Given the difficulties scholars of intercollegiate athletics often face in financing large-scale studies on athletics issues, gaining access to student-athletes on many campuses and even knowing what lines of research are already actively under study at the NCAA, Coakley suggested that "the first and most important strategy for stimulating research on intercollegiate sports is to institutionalize the dissemination of information about the athletic department, sport teams, and athletes to the faculty…

The authors are with the National Collegiate Athletic Association, 700 W. Washington Street, P.O. Box 6222, Indianapolis, IN 46206-6222.

this strategy should be planned in ways that provide research faculty with some form of access to data collected by the NCAA" (Coakley, 2008, p. 19).

Coakley's paper and the resulting discussions on the barriers to conducting high-quality research on intercollegiate athletics were not lost on NCAA president, Myles Brand. Subsequent to the Scholarly Colloquium, Brand charged NCAA senior vice president, Bernard Franklin, and the NCAA research staff with developing a plan for a phased release of previously collected data beginning in 2009. The framework of this plan was announced by Franklin at the second Scholarly Colloquium in January, 2009. The complexities of enacting a data-sharing process at the NCAA are certainly clear at this point. As others have described (e.g., Abbott, 2007; Breckler, 2009), the primary concern is protecting the confidentiality of individuals and institutions that have provided data to the NCAA. It is vital to its membership (as it is within any research setting) that the NCAA adhere to all ethical and legal commitments as data are released in a public manner. As these commitments or confidentiality agreements vary in sometimes subtle ways from study to study, this alone is a difficult process. It is imperative that any data made openly available not only have obvious identifying fields removed, but also be fully deidentified to the rigorous standards necessary to ensure that a sophisticated user could not combine seemingly innocuous data elements to reveal identities or related confidential information. At the same time, the data need to be useful to serious researchers. It turns out that striking a workable balance between protecting research participants and creating useful data for researchers is truly a daunting task.

To assist with these many difficult issues, the NCAA called on a number of experts in social science data sharing. The primary advisor in this process was Dr. Margaret (Maggie) Levenstein, the Executive Director of the Michigan Census Research Data Center. Dr. Levenstein has significant experience in dealing with the complexities of sharing sensitive data, and she skillfully designed the general parameters of the program that the NCAA will put in place. Levenstein also received guidance from the members of the NCAA Data Analysis Research Network (an advisory panel of about 20 scholars in education, testing and psychology chaired by James Jackson, who is a Professor of Psychology at the University of Michigan and the Director of the Institute for Social Research), the NCAA Research Committee (NCAA research oversight body that includes faculty, university administrators and athletics personnel, currently chaired by Kurt Beron, a Professor of Economics at the University of Texas at Dallas), and various members of the NCAA staff.

The NCAA plan includes two important initial steps. The first is the development of a partnership between the NCAA and the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. ICPSR is the largest repository of social science data in the world, housing more than 60,000 distinct data sets. Having access to the expertise and infrastructure of ICPSR will allow the NCAA to overcome substantial structural impediments to developing a comprehensive data sharing program. In particular, a data delivery system can be activated much more quickly (and at much lower cost) than if a native system was developed internally. ICPSR will also make the data available in several standard formats and assist the NCAA with creating proper supporting materials (for example, data codebooks).

The second step in the NCAA plan is to establish a disclosure review committee made up of experts in the field of database management and data sharing, along with

representatives from the NCAA staff and membership. This group will be charged with reviewing NCAA data archives to ensure that all legal, ethical and confidentiality obligations are being met prior to making any data publically available. Additionally, this group will provide advice to the NCAA research staff on methods that will allow for a smooth transition between the collection of future data and their being made available to outside researchers.

Once these initial steps have been taken, the NCAA's goal is to prepare four data collections for public release within the next year. These include a user-friendly, longitudinal graduation-rates database for all colleges and universities in NCAA Divisions I and II; a longitudinal database of team-level Academic Progress Rates (APR) in Division I (APR is a real-time measure of student-athlete academic success as measured by academic eligibility for competition and retention); and individual-level data from the NCAA Study of College Outcomes and Recent Experiences (SCORE) and the Growth, Opportunities, Aspirations and Learning of Students in college (GOALS) study.

The first two of these data collections (graduation rates and APR) consist of team-level information that is partially available to the public currently (although not in a format easily conducive to analysis). It is expected that a number of enhancements will be made to the data before public distribution, including the presentation of single-year data (rather than four-year rolling averages) for all years in which data have been collected and can be made available (currently 7 years for graduation rates and 5 years for APR) and more nuanced data underlying the rate calculations (for example, team-level retention and academic eligibility rates used to calculate APR). The NCAA plans to update these datasets yearly.

The second two data collections (GOALS and SCORE) are individual-level survey data that describe student-athlete perceptions of their college experiences. The GOALS survey covered a wide range of topics including academic, athletics and social experiences, health, time demands, and general well-being. Approximately 20,000 student-athletes across NCAA Divisions I, II and III participated in the GOALS study while they were in college in 2007. The SCORE study assessed similar topics among thousands of former high school and college student-athletes (primarily those recruited by or participating in sports at Division I schools) who were surveyed 11 years after leaving high school. The SCORE survey additionally examined the educational trajectories of these former student-athletes and assessed characteristics of their current employment. These data collections have been particularly valuable for NCAA staff and committees and should prove interesting to scholars who wish to study the attitudes and behaviors of college student-athletes.

The sequential publication of these four data collections is expected to allow the NCAA to hone its ability to deliver useful data to researchers of intercollegiate sport while developing protocols to ensure the protection of research participants. All advisors in this process were clear that it is important to release these databases deliberately so that the NCAA can learn about and correct unanticipated problems quickly and efficiently. In the long run, the NCAA expects to follow this initial release with a sharing of as much data as possible from its archives. As scholars in various disciplines gain access to NCAA data, we hope to gain a better understanding of exactly which studies are of greatest interest to researchers and adjust data sharing plans accordingly. However, the breadth and timing of the NCAA's data sharing process will

be constrained by staff resources and the complexity of the issues surrounding each dataset. Some NCAA data may be made accessible to the general public, while other databases may require a more restrictive process be put in place (e.g. restricted-use data agreements, or analysis within an enclave setting). Again, protection of research participants has to remain the key element for the NCAA in all decisions on whether or how to share research data.

We have been told to expect headaches and additional hurdles along the way to implementing a comprehensive data sharing program. But, we also expect that efforts in establishing and maintaining a data sharing program will be offset by a greater good of enhancing research that will directly benefit students, colleges and intercollegiate sport. We also hope this initiative will enhance the dialogue between NCAA research staff and outside scholars. Data-driven policy analysis has become a key aspect of decision making among NCAA member schools and leaders; broader access to NCAA data by others interested in education and sport can only improve these efforts.

References

Abbott, A. (2007). Notes on replication. Sociological Methods & Research, 36(2), 210–219.

Azar, B. (1999). Will you be forced to share your data? [Electronic version]. Monitor on Psychology, 30(8). Retrieved February 27, 2009, from http://www.apa.org/monitor/sep99/data.html.

Breckler, S.J. (2009). Dealing with data. Monitor on Psychology, 40(2), p. 41.

Coakley, J. (2008). Studying intercollegiate sports: High stakes, low rewards. Journal of Intercollegiate Sports, 1(1), 14–28.

DeAngelis, T. (2004). Data sharing: A different animal. Monitor on Psychology, 35(2), p. 48.

Freese, J. (2007a). Overcoming objections to open-source social science. Sociological Methods & Research, 36(2), 220–226.

Freese, J. (2007b). Replication standards for quantitative social science. Sociological Methods & Research, 36(2), 153–172.

Gleditsch, N.P., Metelits, C., & Strand, H. (2003). Posting your data: Will you be scooped or will you be famous? International Studies Perspectives, 4, 89–97.

King, G. (2007). An introduction to the Dataverse Network for data sharing. Sociological Methods & Research, 36(2), 173–199.