ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

New Orleans JAZZED about Engineering Education

Paper ID #15136

Expanding Access to and Participation in the Multiple Institution Database for Investigating Engineering Longitudinal Development

Dr. Matthew W. Ohland, Purdue University, West Lafayette

Matthew W. Ohland is Professor of Engineering Education at Purdue University. He has degrees from Swarthmore College, Rensselaer Polytechnic Institute, and the University of Florida. His research on the longitudinal study of engineering students, team assignment, peer evaluation, and active and collaborative teaching methods has been supported by the National Science Foundation and the Sloan Foundation and his team received Best Paper awards from the Journal of Engineering Education in 2008 and 2011 and from the IEEE Transactions on Education in 2011. Dr. Ohland is Chair of the IEEE Curriculum and Pedagogy Committee and an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE and IEEE.

Mr. Russell Andrew Long, Purdue University, West Lafayette

Russell Long, M.Ed. is Director of Project Assessment at the Purdue University School of Engineering Education and Managing Director of The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD). He has extensive experience in performance funding, large data set analysis, program review, assessment and student services in higher education. One of his greatest strengths lies in analyzing data related to student learning outcomes and, therefore, to improving institutional effectiveness. His work with MIDFIELD includes research on obstacles students face that interfere with degree completion and, as well, how institutional policies affect degree programs. His group's work on transfer students, grade inflation, and issues faced across gender and ethnicity have caused institutions to change policies so that they may improve. Awards and publications may be found at https://engineering.purdue.edu/people/russell.a.long.1.

Dr. Susan M Lord, University of San Diego

Susan M. Lord received a B.S. from Cornell University and the M.S. and Ph.D. from Stanford University. She is currently Professor and Chair of Electrical Engineering at the University of San Diego. Her teaching and research interests include electronics, optoelectronics, materials science, first year engineering courses, feminist and liberative pedagogies, engineering student persistence, and student autonomy. Her research has been sponsored by the National Science Foundation (NSF). Dr. Lord is a fellow of the ASEE and IEEE and is active in the engineering education community including serving as General Co-Chair of the 2006 Frontiers in Education (FIE) Conference, on the FIE Steering Committee, and as President of the IEEE Education Society for 2009-2010. She is an Associate Editor of the IEEE Transactions on Education. She and her coauthors were awarded the 2011 Wickenden Award for the best paper in the Journal of Engineering Education and the 2011 Best Paper Award for the IEEE Transactions on Education. In Spring 2012, Dr. Lord spent a sabbatical at Southeast University in Nanjing, China teaching and doing research.

Dr. Marisa K. Orr, Louisiana Tech University

Dr. Orr is an Assistant Professor in Mechanical Engineering and Associate Director of the Integrated STEM Education Research Center (ISERC) at Louisiana Tech University. She completed her B.S., M.S., and Ph.D. in Mechanical Engineering, as well as a Certificate of Engineering and Science Education at Clemson University. Her research interests include student persistence and pathways in engineering, gender equity, diversity, and academic policy.

Dr. Catherine E. Brawner, Research Triangle Educational Consultants

Catherine E. Brawner is President of Research Triangle Educational Consultants. She received her Ph.D.in Educational Research and Policy Analysis from NC State University in 1996. She also has an MBA from Indiana University (Bloomington) and a bachelor's degree from Duke University. She specializes in evaluation and research in engineering education, computer science education, teacher education, and

technology education. Dr. Brawner is a founding member and former treasurer of Research Triangle Park Evaluators, an American Evaluation Association affiliate organization and is a member of the American Educational Research Association and American Evaluation Association, in addition to ASEE. Dr. Brawner is also an Extension Services Consultant for the National Center for Women in Information Technology (NCWIT) and, in that role, advises computer science departments on diversifying their undergraduate student population. Dr. Brawner previously served as principal evaluator of the NSF-sponsored SUCCEED Coalition. She remains an active researcher with MIDFIELD, studying gender issues, transfers, and matriculation models in engineering.

Expanding Access and Participation in the Multiple Institution Database for Investigating Engineering Longitudinal Development

Retention and graduation are the dominant metrics for studying student success in engineering education and in higher education in general; yet available national datasets do not facilitate establishing national retention/graduation benchmarks. A national, longitudinal, engineering student unit-record database would make it possible to calculate retention and other metrics consistently. This would permit benchmarking, support peer comparisons, and the use of new metrics backed by community support.

Sharing longitudinal student record data is critical to addressing important questions that are being asked of higher education. The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) is a multi-institution, longitudinal, student record level dataset that is used to answer many research questions about how students maneuver through required engineering curricula and what obstacles stand in their way towards graduation. MIDFIELD comprises whole population unit-record data for undergraduate, degree-seeking students—including students who matriculate in engineering, those who migrate into engineering from other majors, students who come to engineering as transfer students, part-time engineering students, and students who have never enrolled in engineering. This diversity results in a dataset that currently comprises twenty-five years of data that includes 1,014,887 unique undergraduate, degree-seeking students. Of those students 210,725 were ever enrolled in engineering. While the original database contains only eleven institutions, the plan for MIDFIELD has always been to expand the database to include all public institutions in the United States that offer undergraduate programs in engineering. An award by the National Science Foundation (#1545667, $4,010,978.00, 03/01/16 to 02/28/2021) will support increasing the number of partner institutions to 103. Students in the expanded MIDFIELD will comprise over half of the undergraduate engineering degrees awarded at U. S. public institutions and approximately two-thirds of the U. S. undergraduate engineering student population in any given year during the past 30 years. The expanded MIDFIELD will contain unit record data for over 10 million individual students and will contain minority serving institutions and institutions from a broad range of research classifications.

Designing a national student unit-record data system

Four design principles have been identified for expanding MIDFIELD into a national unit-record database, based on input from interviews and focus groups with engineering administrators, engineering education researchers, registrars, institutional research staff, and data archivists..

Data should be accessible to a broader community of researchers. Institutional representatives that were interviewed recognized the benefits of allowing researchers to have access to a national student unit-record data system. In addition to accelerating current research, permitting access to a broader research community would attract the research interest of demographers, sociologists, statisticians, and others to research questions of interest to engineering education.

Partner institutions must not be affected negatively by published research results. To protect partner institutions, names of MIDFIELD partners should not be associated publicly with specific statistics or calculations. Tables and figures displaying results should mask the identities of institutions in the data. Institution names should be used only when data is aggregated across more than one institution, and only then so long it is not possible to deduce the institutions.

Partner institutions should have special access to conduct peer comparisons. Institutional representatives were clearly interested in the opportunity to use MIDFIELD data to conduct peer comparisons in greater detail than they have access to with the data available currently. At the same time, they were unwilling to allow other institutions to have that level of access to their data without some indication of shared risk and trust. Further, findings from such studies should not have the opportunity to have a negative effect on institutions. The results from such peer comparisons must be used solely for institutional analysis and only information pertaining to the institution itself may be made public.

All institutions should have equal access to benefit from the MIDFIELD partnership. To ensure that MIDFIELD does not become a resource that further privileges schools that have the resources to participate, we must find resources for institutions to extract the historical data needed join the MIDFIELD partnership. Yet, admission to the partnership is not sufficient to level the playing field. Well-resourced institutions are more likely to have highly skilled researchers who conduct research and publish findings based on MIDFIELD. This benefit cannot be granted to MIDFIELD partners, but a corollary benefit can be assured – that less-resourced institutional partners benefit when other institutions conduct research using MIDFIELD. For this reason, while published research that generates institutional findings must mask institutional identity, institutions must privately be informed of their own identity. Thus, researchers at all institutions using MIDFIELD provide an institutional research benefit to all the MIDFIELD partners.

A valuable partnership in data sharing. The Interuniversity Consortium for Political and Social Research specializes in handling and sharing large datasets. In partnership with ICPSR, the authors have negotiated a complex restricted-use data dissemination agreement that describes a process by which MIDFIELD partner institutions provide institutional data. MIDFIELD staff convert the institutional data to the MIDFIELD common format and transmit the common format data to ICPSR.ICPSR archives the data, administers and enforces data use agreements, and provides access to the data to investigators who have executed data use agreements. Two distinct data use agreements implement these requirements: a "restricted data use agreement for research" and a "restricted data use agreement for institutional analysis". Signatures are being sought from the current partners, and all institutions that join the partnership in the future will be expected to participate in this archive. ICPSR will control the distribution of archived data and will manage risk through restricted-use data dissemination agreements. MIDFIELD staff will continue to add institutions to the archive as agreements are reached with MIDFIELD partners. Derived variables will be added to the common format during updates. MIDFIELD staff will distribute a smaller "dummy" data file with valid variable values for use in workshops and by researchers who want to explore MIDFIELD before contracting with ICPSR to gain access.

A timeline for expansion of institutional partners and research access

The expansion of institutional participation is limited by trust, politics, and other factors. It is unrealistic to expect that MIDFIELD will ever include data from all the U.S. institutions with baccalaureate programs accredited by the Engineering Accreditation Commission of ABET. Research access to the MIDFIELD dataset is limited by concerns for institutional and individual privacy and the liabilities related to those. In spite of these constraints, there are plans to expand both the number of participating institutions and research access to the dataset.

Expansion strategy. New institutional partners will receive funding to provide and update data. As the database becomes larger in size, joining the MIDFIELD partnership becomes even more attractive. Twenty institutions have signed letters of commitment to join MIDFIELD. New institutions will be targeted to reflect variability in geographic region, institution size as determined by the number of engineering graduates per year, and institutional control (public or private). Institutions will also be targeted that have a high or low graduation rate for under-represented minorities – plans include adding 5 Historically Black Colleges and Universities (HBCUs), 7 Hispanic Serving Institutions (HSIs), 5 institutions with high Native American populations and 7 universities with high Asian/Pacific Islander populations. Including the current MIDFIELD institutions (11 public institutions with 9 in the Southeast, 1 in the Midwest, and 1 in the West), the expanded MIDFIELD will include the following types of institutions:

By region:
Northeast – 13 private, 11 public
Southeast – 7 private, 23 public
Midwest – 6 private, 12 public
Southwest – 2 private, 9 public
West – 6 private, 14 public

By number of engineering graduates:
Fewer than 300 graduates – 20 private, 21 public
301 to 500 graduates – 10 private, 14 public
501 to 1,000 graduates – 4 private, 18 public
Greater than 1,001 graduates – 16 public

The collection of institutional data will proceed in seven phases (see Figure 1) – each year adding approximately 20 institutions. Multiple activities occur in each phase.

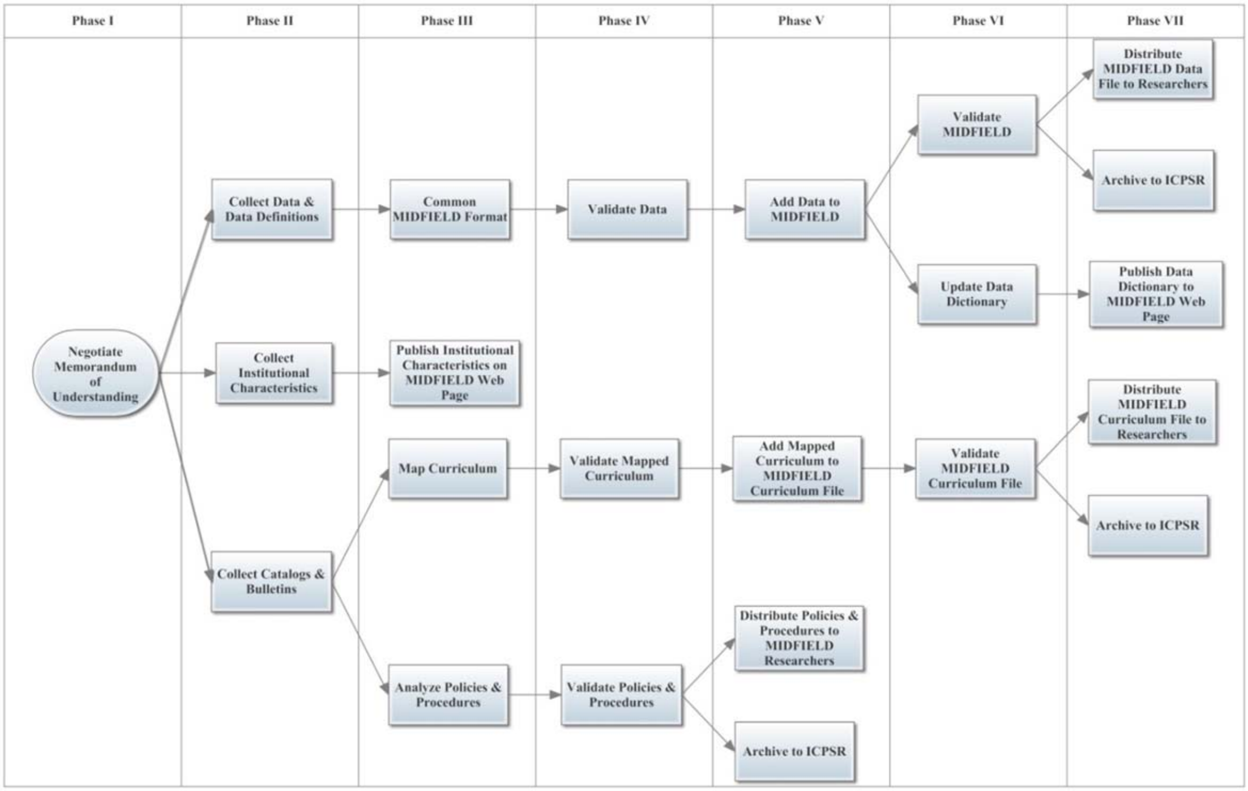
Phase I
Phase II
Phase III
Phase IV
Phase V
Phase VI
Phase VII
Negotiate Memorandum of Understanding
Collect Data & Data Definitions
Common MIDFIELD Format
Validate Data
Add Data to MIDFIELD
Validate MIDFIELD
Distribute MIDFIELD Data File to Researchers
Archive to ICPSR
Update Data Dictionary
Publish Data Dictionary to MIDFIELD Web Page
Collect Institutional Characteristics
Publish Institutional Characteristics on MIDFIELD Web Page
Map Curriculum
Validate Mapped Curriculum
Add Mapped Curriculum to MIDFIELD Curriculum File
Validate MIDFIELD Curriculum File
Distribute MIDFIELD Curriculum File to Researchers
Archive to ICPSR
Collect Catalogs & Bulletins
Analyze Policies & Procedures
Validate Policies & Procedures
Distribute Policies & Procedures to MIDFIELD Researchers
Archive to ICPSR

Figure 1: MIDFIELD Data Collection Phases

Succession plan. Along with plans for expansion, a succession plan is being developed for both the MIDFIELD project director and the data steward. As new institutional partners are added to MIDFIELD, some of those new relationships build on existing relationships, but some prospective partners have already approached the MIDFIELD team about joining the project. This is a sign that MIDFIELD researchers have earned the trust of the community through the quality of their work, by the rigorous protection of student and institutional confidentiality, and by respect for the trust that has already been extended by other institutions through the release of student data.

Expansion of research access. Archiving the dataset with ICPSR represents an important long-term solution to expanding research access to MIDFIELD, and institutions that join the MIDFIELD partnership are asked to commit to participating in that archiving process. Researchers who do gain access to the MIDFIELD data must sign a confidentiality agreement that specifies the terms of use of MIDFIELD data.

Data Security

The computers on which MIDFIELD data reside are not connected to the internet. Member institutions transmit data to the MIDFIELD data steward via password-protected, encrypted files. Physical files are stored in a locked filing cabinet in a secure office. Only the MIDFIELD data steward and project director have access to these files. Student identifiers are created especially for MIDFIELD – they are not Social Security Numbers or student IDs. MIDFIELD data is cleaned, verified, and backed up weekly.

Student Confidentiality

Data security is only the beginning of the protection of student data. Ironically, the very fact that MIDFIELD has student records for over one million students makes it easier to protect the confidentiality of individual students—their identity is protected by primarily by reporting only aggregated results. While research using MIDFIELD conforms to standard cell-size limitations in its research designs, the large population in MIDFIELD frequently permits the adoption of stricter minimum cell sizes that both protect students and give greater confidence in the results. Furthermore, MIDFIELD researchers avoid reporting too much information about groups of students. For example, when discussing outcomes of a population that is disaggregated by race/ethnicity, gender, and discipline, aggregating those students across multiple institutions researchers can provide protection for both students and institutions. Researchers are discouraged from using MIDFIELD data to predict the behavior or outcomes of an individual. MIDFIELD cannot predict what a student will do. MIDFIELD is best used to reveal what large numbers of students have done and suggest what others who share certain characteristics might do.

Institutional Confidentiality

To avoid harming the institutional partners and MIDFIELD's relationship with them, validation of both the MIDFIELD dataset and any results released publicly is critical. This has resulted in a long learning curve for new researchers in developing both the expertise and the confidence to publish results. The challenges extend well beyond knowledge of data management and statistical procedures in a general sense. The primary challenge lies in MIDFIELD-specific issues of merging data from institutions that have different data handling practices, different data schema, different academic policies, and different institutional histories. Once the data are in the MIDFIELD common format, many differences in data-handling practices have been smoothed over, yet some remain. Students who are planning to pursue engineering but have yet not selected a specific discipline are tracked in ways that vary by institution, when the student expresses interest in engineering, and whether they meet engineering's admission standards. How this is sorted out can affect reported matriculation patterns and retention rates. Institutions track participation in cooperative education in different ways, so there is a difference between whether a student is on co-op in a particular term (in the term table) and whether a student has ever participated in the co-op program (a logistic variable in the demographic table). Each institution has policies regarding academic probation, suspension, expulsion, and readmission, but the criteria defining each of those differ as do each institution's related supports and consequences for students. Some of the biggest challenges are those of institutional context, because those have been learned over a long period of partnership with the institutions and through more intensive interviews of knowledgeable personnel at institutions joining the partnership more recently. Researchers might be surprised that no students graduated from Georgia Tech in Summer 1996, until they are reminded that Atlanta hosted the Olympic and Paralympic Games and that the Olympic Village that housed visiting athletes occupied most of the Georgia Tech campus. The learning curve involved in using MIDFIELD safely and effectively is an ongoing challenge for the project, particularly as we seek to share MIDFIELD data with an ever-larger community of researchers, including those with whom we might not have direct contact.

MIDFIELD avoids linking findings to specific institutions. Many of the methodological approaches used to conceal institutional identity are simple—reporting percentages to mask institution size, using a separate institutional key for multiple graphs in the same publication to avoid the cumulative loss of anonymity, and aggregating data across institutions. While institutional policies are typically public, findings are only linked to policies where institutional confidentiality was protected (because those policies were common to multiple institutions). We strive to explore institutional variability without compromising three important principles:

1. Institutional data are provided to the MIDFIELD project on the condition that researchers protect the identity of the partner institutions and each institution's students.
2. Increasingly specific institutional descriptions discourage readers from considering MIDFIELD research to be generalizable, in spite of other significant evidence that there is much that is common among engineering programs and their interaction with students.
3. While MIDFIELD includes data for very large numbers of students, a relatively small number of institutions are represented, so institutional variation must be treated using a case study approach. Conscientious institution-level analysis would require a large number of diverse institutions.

This last principle can be difficult in cases where it appears that something is specific to an institution, a particular type of institution (HBCU), or institutions that share a particular policy. When researchers (including the authors) begin to speculate along those lines, others on the research team are expected to recall one of the team's catchphrases: "If the institution is the unit of analysis, we only have a sample size of 11." This limitation affects studies such as comparing outcomes of first-year engineering programs with those of institutions where students matriculate directly to a discipline.

The benefits of a large-scale student unit-record data system

It might seem that the greatest benefit of adding more institutions to MIDFIELD would be to make the database more representative of the set of U.S. institutions offering B.S. degrees in engineering and to make the findings more generalizable. Of greater interest is creating the conditions to answer research questions that require or would benefit from an institutional unit of analysis. Such studies fall into several categories:

Studies of academic policies. Academic policies certainly affect the educational environment. Adding institutions to MIDFIELD would allow researchers to establish clearer links between those policies and the educational outcomes of students.

Studies of curricular structure. To conduct a robust study of the influence of curricular structure, the database must include not only a larger number of institutions, but institutions representing a greater diversity of curricular models.

Studies that depend on institution-level variables. Such studies can measure the influence of institution size, engineering fraction of enrollment, private vs. public control, and variables related to financial need. While these have been studied using other datasets, there is much to learn from studying these in multilevel models including institution and student-unit-record data.

Multiple strategies are needed to engage a broad community in educational research

We propose to help prepare a community of researchers to engage in high-quality data-intensive research. A free 90-minute workshop will reach a wide audience, communicate the benefits and challenges of using MIDFIELD, and make participants aware of the need for better data displays. The central component is a two-day, self-sustaining MIDFIELD Institute that develops the skills to analyze and display complex educational data.

Reaching a wide audience with a 90-minute workshop

Well-designed, interactive workshops are an established technique for promoting effective practice. The MIDFIELD workshop—*Learning from Longitudinal Student Data*—will engage participants, solicit their opinions and experiences, connect their stories to research findings, build a common vocabulary, provide training in relevant methods, and challenge them to change their practice. Workshop participants will use a mock dataset of student unit-record data in the MIDFIELD format, but with no actual student data. The workshop will be offered regularly at conferences attracting researchers from target disciplines such as the American Society for Engineering Education Annual Conference, Frontiers in Education, EDUCON, the Annual Meetings of the American Educational Research Association (AERA), and the Association for Institutional Research (AIR) Forum. Some participants will continue on to the MIDFIELD Institute.

The workshop is interactive and lasts 90 minutes—both are commonly requested workshop attributes. The scope and length of the agenda (below) will be adjusted to the needs and resources of the audience and venue. At the conclusion of the workshop, participants should be able to: (1) list common elements of student unit-record data, (2) list and define derived variables available in MIDFIELD, (3) calculate and evaluate educational metrics using MS Excel, (4) identify deficiencies of common graph types.

The MIDFIELD Institute to promote high-quality data-intensive educational research

A multi-day MIDFIELD Institute will be developed and held each January at the University of San Diego (USD). While NSF funds will support the development of the Institute during its first few years, NSF funds will not support hosting ongoing delivery of the Institute. Registration fees will cover ongoing cost of facilities, meals, facilitators, and administrative support.

The Institute uses the same MIDFIELD-format mock dataset created for the workshop. Participants will engage in guided practice in multiple forms of data exploration and then explore research questions of their own. Facilitators will guide participants to produce data displays in Excel that are based on contemporary principles of perception, rhetoric, and design. Participants will share data displays related to their own research questions to engage in iterative data exploration and communication and receive feedback from peers and workshop facilitators. Participants will learn how other data sources have been merged with MIDFIELD, such as IPEDS, NCES, ASEE, NCEES, NSSE, and US Census data. We will share how qualitative data inform causality and the search for deeper meaning. Participants will share their results through a conference special session to encourage others to a higher level of engagement in data-intensive educational research.

To encourage a diverse community of researchers to attend the MIDFIELD Institute, travel stipends are budgeted for up to 10 attendees per year from Minority Serving Institutions. Each Institute will serve 20 participants to ensure high-quality interaction and hands-on use of the dataset. The Institute provides a training model including information, demonstration, practice, and feedback according to the agenda that follows.

MIDFIELD Institute learning objectives and agenda

At the conclusion of the Institute, participants should be able to: (1) describe data available in MIDFIELD, (2) calculate and evaluate educational metrics using MS Excel, (3) identify deficiencies of common graph types, (4) describe key principles of effective data visualization, (5) use MIDFIELD to produce a table of data that addresses a research question of their choosing, (6) use Excel to explore and tell a story from their data, (7) list potential resources beyond MIDFIELD that can contribute to answering their research questions.

Educating the research community about the display of quantitative data

Data visualization is more than creating good graphs. It is an important step in an iterative process of discovering and presenting the story data have to tell. To illustrate our approach, we show below a published graph and describe how to apply contemporary design principles to create a more effective graph. The "Gains in Retention" graph, Figure 2, is displayed in a way that distorts the message. The nearly-constant 8–10% difference between comparison groups is exaggerated because the displayed difference is always more than 10% of the graph scale. Moreover, the retention difference between the two programs appears more significant as time passes: in semester 5, the 10% difference is visually represented by a ratio of bar heights of 1.8 to 1; in semester 7, by a ratio of 3.6 to 1—over three times as large as the real difference. The second main deficiency is the omission of the first semester, when 100% of both populations are enrolled. The original graph fails to visually display that greatest change in the two populations occurs between the first and third semesters. Thereafter, the two populations differ by a nearly constant amount. These deficiencies are addressed in our redesign in Figure 3.

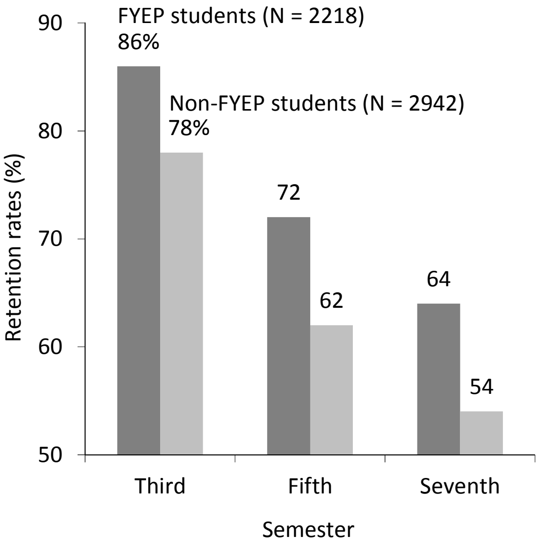

Figure 2. Reproduction of the published design.

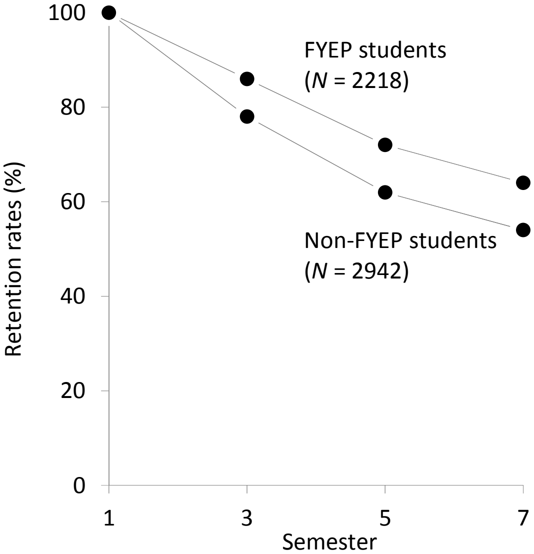

Figure 3. A redesigned graph tells a clearer story.

The expanded MIDFIELD will be an essential tool for institutional researchers to study students on the local, regional, and national level. Broader access to MIDFIELD data through a data archive will leverage the investment in its infrastructure and increase the diversity and pace of research using the database. Expanding access to MIDFIELD should result in the development of a research community that shares best practices for using this data, leading to methodological advances as well. Adding new institutional partners will enhance the generalizability of this research and allow a larger community of researchers to access this resource resulting in a dramatic increase in high-quality research.