A Primer on Working with Longitudinal Student Unit Records

Mr. Russell Andrew Long, Purdue University

Russell Long, M.Ed. was the Director of Project Assessment at the Purdue University School of Engineering Education (retired) and is Managing Director of The Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD).

Richard A. Layton, Layton Data Display

Richard A. Layton is Professor Emeritus of Mechanical Engineering at Rose-Hulman Institute of Technology. He received a B.S. from California State University, Northridge, and an M.S. and Ph.D. from the University of Washington. With Matthew Ohland, Layton is a co-founding developer of the CATME Smarter Teamwork system and the midfieldr R package for working with student unit records. He is a co-author of the Engineering Communications Manual, Oxford Univ. Press, 2017. He currently consults as a data visualization specialist using R.

Dr. Marisa K. Orr, Clemson University

Marisa K. Orr is an Associate Professor in Engineering and Science Education with a joint appointment in the Department of Mechanical Engineering at Clemson University.

Dr. Susan M. Lord, University of San Diego

Susan Lord is Professor and Chair of Integrated Engineering at the University of San Diego. She received a BS from Cornell University in Materials Science and Electrical Engineering (EE) and MS and PhD in EE from Stanford University. Her research focuses on the study and promotion of equity in engineering including student pathways and inclusive teaching. She has won best paper awards from the Journal of Engineering Education, IEEE Transactions on Education, and Education Sciences. Dr. Lord is a Fellow of the IEEE and ASEE and received the 2018 IEEE Undergraduate Teaching Award. She is a coauthor of The Borderlands of Education: Latinas in Engineering. She is a co-Director of the National Effective Teaching Institute (NETI).

Dr. Matthew W. Ohland, Purdue University

Matthew W. Ohland is the Dale and Suzi Gallagher Professor and Associate Head 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 and forming and managing teams has been supported by the National Science Foundation and the Sloan Foundation and his team received for the best paper published in the Journal of Engineering Education in 2008, 2011, and 2019 and from the IEEE Transactions on Education in 2011 and 2015. Dr. Ohland is an ABET Program Evaluator for ASEE. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS.

A Primer on Working with Longitudinal Data

Abstract

Longitudinal, student-level data are a rich resource for characterizing how students navigate the terrain of higher education. Learning to work effectively with such data, however, can be a challenge. In this paper, we share some of our experiences over years of conducting research with the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD). MIDFIELD contains individual student-level records for all undergraduate students at 19 US institutions with over 1.7 million unique students. This paper focuses on our lessons learned about processing longitudinal data to prepare it for analysis. We describe and define the steps that we take to process the data including filtering for data sufficiency, degree-seeking, and program (major), then classifying by completion status and demographics. We use the examples of calculation of graduation rate and stickiness to show the details of how the processed data is used in analysis. We hope this paper will help introduce the landscape of longitudinal research to a wider audience and provide tips for working with this valuable resource.

Introduction

The study of engineering education has been enhanced by the creation and study of a multi-institution student records longitudinal dataset. Longitudinal data contains the same variables for the same individual over time. Longitudinal student records data is powerful but learning to work with it can be daunting. In this paper, we share some of our experiences over many years of conducting research with the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [1]. MIDFIELD contains student record data for all undergraduate students at 19 institutions across the USA with over 1.7 million unique students. This rich dataset is large enough to permit disaggregation by multiple categories such as race/ethnicity, sex, and program. Such disaggregation is particularly important for conducting intersectional analyses and investigating small, underrepresented populations. This has enabled impactful and award-winning research and informed institutional decision-making. For examples, see [2, 3, 4].

Before data analysis can begin, researchers need to process the dataset to define what groups of students will be studied and what additional variables need to be created. This paper describes and defines the steps that we take to process the data. Our initial data processing involves eight steps: 1) measuring data sufficiency, 2) determining if a student is degree-seeking, 3) determining program (major), 4) creating blocs, 5) setting first-year engineering (FYE) proxies, 6) determining starters in each major, 7) identifying students who graduate, and 8) creating grouping variables. We also summarize two different analysis metrics: graduation rate and stickiness. The following discussions are focused on MIDFIELD data, but the concepts can be applied to any longitudinal dataset.

For example, if researchers want to study the graduation rate of students in electrical engineering (EE) disaggregated by race/ethnicity and sex, they will first choose only students that started in electrical engineering, as defined by their Classification of Instructional Programs (CIP) code

[5], that had data sufficiency. Data sufficiency means that they have enough data to compute a 6-year graduation rate. Some student records near the lower and upper terms that bound the available longitudinal data must be excluded to prevent false summaries involving timely completion. Timely completion is the count of graduates completing their programs in no more than 6 years. Then the researcher would choose only degree-seeking students. It may be helpful to create variables with thoughtfully decided names along the way such as the name of the degree program. For example, there might be a variable for students ever in EE and another for those who graduate in EE. Before applying any metrics, data must be filtered for the demographic variables of interest which can include race/ethnicity and sex. Data must be grouped and summarized.

See Appendix 1 for definitions of the terms that we use in this paper.

Step One: Measure Data Sufficiency

The time span (or range) of MIDFIELD data varies by institution. At the upper and lower limits of a data range, a potential for false counts exists when a metric (such as graduation rate) requires knowledge of timely degree completion. For such metrics, student records that produce problematic results due to insufficient data are nearly always excluded from study. See [6] for the R code that we use to determine data sufficiency.

Upper-limit data sufficiency

For students admitted too near the upper limit of their institution's data range, the available data cover an insufficient number of years to know if completion is timely. To illustrate, in Figure 1 we compare two students admitted in different terms with representative time spans shown for timely completion. In this scenario, we assume institution data is available from 1986 to 1996.

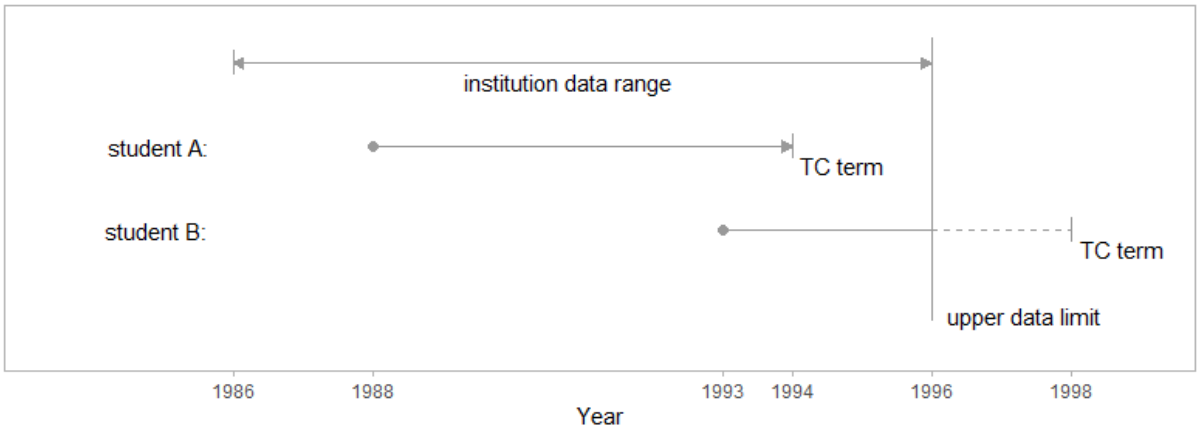

Figure 1: Upper limit data sufficiency.

Student A

Student A enters in 1988 with a timely completion (TC) term in 1994. In both of the following cases, the data sufficiency criterion is satisfied since the TC term is within the range of the available data, and the records are included in a study.

- A-1: First time in college (FTIC), so we know their first term is their entry term (i.e., they are not a continuing student) and we can determine their TC term.
- A-2: Transfer student, and we know their first term in a MIDFIELD institution. We have no knowledge of how much time was spent accumulating their pre-MIDFIELD credit hours, but we can estimate a TC term with respect to their "level" at entry, that is, entering as a first-year student, second-year student, etc.

Student B

Student B enters in 1993 with a TC term in 1998, two years beyond the range of the data. We have several possible cases,

- B-1: Before the upper data limit, the student completes their program (timely completion, known record)
- B-2: Before the upper data limit, the student leaves the data base (non-completion, known record)
- B-3: After the upper data limit, the student completes before their TC term (timely completion, no record)
- B-4: After the upper data limit, the student completes after their TC term or fails to complete (late completion or non-completion, no record)

Because the outcomes in cases B-3 and B-4 are not in the record, to include case B-1 and B-2 produces a miscount of timely completers, late completers, and non-completers. Thus, all student B records are excluded from the study.

Lower-limit data sufficiency

To determine data sufficiency record exclusions at the lower limit of the data range, we compare a student's first term (non-summer) to the first term of the data range (also non-summer). When these two terms are identical, the complete unit record is excluded. We illustrate with the three scenarios described below.

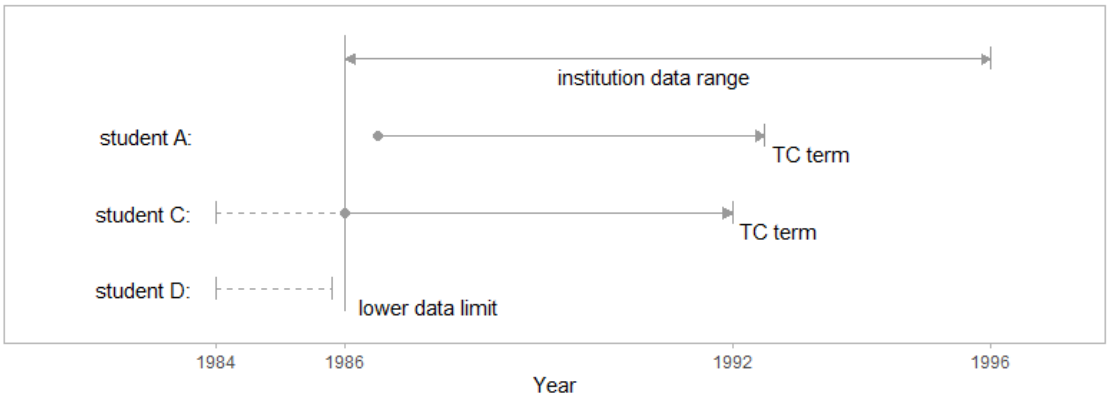

Figure 2: Lower limit data sufficiency.

Student A

Like Student A in Figure 1, they enter the dataset in a term following the data lower limit and are included in a study.

Student C

Student C enters the institution before the lower limit of the data range (a "continuing" student) or they enter the institution at the lower limit precisely.

- C-1: If student C is continuing, regardless of status (FTIC or transfer), making an estimate of their TC term invariably leads to false counts because we have no knowledge of how much time was spent accumulating credit hours at their MIDFIELD institution before the lower data limit. Including C-1 would also produce false counts because of student D (discussed below).
- C-2: If student C is not continuing, that is, their first-time entry to a MIDFIELD institution is at the lower data limit (here, 1986), we would include them in a study if we could. Unfortunately, we cannot distinguish them from continuing students. Having to exclude C-1 inherently excludes C-2 as well.

Student D

Student D enters the institution at the same time as continuing student C but leaves the database before the data lower limit term.

- D-1: Student D did not timely-complete their program. In this case, if we include student C our count of *non-completers* is low (D-1 cases are missing), resulting in an inflated ratio of completers to non-completers.
- D-2: Student D did timely-complete their program. Here, if we include student C our count of *completers* is low (D-2 cases are missing), resulting in a diminished ratio of completers to non-completers.

The balance of these two effects is unknowable. Since student D cannot possibly be included, Student C must also be excluded.

Step Two: Determine if a student is degree-seeking

Most analysis of student-level records omit records of students not seeking degrees. By design, MIDFIELD contains records of degree-seeking students only. If your dataset contains non-degree students, see [7] for the R code that we use to determine if a student is degree -seeking.

Step Three: Academic programs

In the USA, instructional programs are encoded by 6-digit numbers curated by the US Department of Education known as the classification of instructional programs or CIP code [5]. The US standard encoding format is a two-digit number followed by a period, followed by a four-digit number, for example, 14.0102. MIDFIELD uses the same numerals, but omits the period, i.e., 140102, and stores the variable as a character string.

Taxonomy

Academic programs have three levels of codes and names:

- 6-digit code, a specific program
- 4-digit code, a group of 6-digit programs of comparable content
- 2-digit code, a grouping of 4-digit groups of related content

Specialties within a discipline are encoded at the 6-digit level, the discipline itself is represented by one or more 4-digit codes (roughly corresponding to an academic department), and a collection of disciplines are represented by one or more 2-digit codes (roughly corresponding to an academic college).

For example, Geotechnical Engineering (140802) is a specialty in Civil Engineering (1408) which is a department in the college of Engineering (14). A 2-digit program can include anywhere from four 4-digit programs (e.g., code 24 Liberal Arts and Sciences, General Studies and Humanities) to 238 4-digit programs (e.g., code 51 Health Professions and Related Clinical Sciences). And 4-digit programs include anywhere from one 6-digit program (e.g., code 4100 above) to 37 6-digit programs (e.g., code 1313 Education).

Unfortunately, some disciplines can comprise more than one 4-digit code. For example, the programs that comprise the broad discipline of Industrial and Systems Engineering encompass four distinct 4-digit codes: 1427 Systems Engineering, 1435 Industrial Engineering, 1436 Manufacturing Engineering, and 1437 Operations Research. Hence the importance of being able to search all CIP data for programs of interest.

See [8] for the R code that we use to determine programs.

Step Four: Blocs

A *bloc* is a grouping of student-level records dealt with as a unit, for example, a grouping of starters in a program, graduates of a program, or ever enrolled in a program. We often use a *left join* merging operation to add one or more variables to a working data frame and filter on those variables to construct the desired bloc.

Different metrics require different blocs. Graduation rate, for example, requires starters and their graduating subset whilestickiness requires ever enrolled and their graduating subset. Because a bloc is usually defined for specific programs, the final filter applied in gathering a bloc is often an *inner join* to filter by program labels, as derived in the Programs section.

See [9] for the R code that we use to create blocs.

Step Five: Determine starters in each major

A degree-seeking student enrolled in their first degree-granting program is a *starter* in that

program. Identifying starters is typically performed as part of a graduation rate calculation, though it can also be a useful measure on its own.

Special cases

In two special cases, an entering student's CIP code does not correspond to a degree-granting program. Our procedure for identifying starters accommodates both special cases.

Case 1 Unspecified

The first case includes records for which a CIP is unspecified or reported as "undecided". In MIDFIELD data, both conditions are encoded as CIP 999999. Students may *enter* with this CIP but we do not consider them *starters* until and if they enroll in a degree-granting program.

Case 2 First-Year Engineering (FYE)

The second case is more nuanced. At some US institutions, engineering students are required to complete a First-Year Engineering (FYE) program as a prerequisite for declaring an engineering major. These students are admitted as Engineering majors, but we don't know to which degree-granting program they intended to transition. At the 2-digit CIP level, FYE students are starters in Engineering (CIP 14). If we do not restrict a study to 2-digit CIPs, however, we use FYE proxies—our estimates of the degree-granting engineering programs (6-digit CIP level) that FYE students would have declared had they not been required to enroll in FYE.

Potential for starter miscounts

To illustrate the potential for miscounting starters, suppose we wish to calculate a Mechanical Engineering (ME) graduation rate. Students starting in ME constitute the starting pool and the fraction of that pool graduating in ME is the graduation rate.

At FYE institutions, an ME program would typically define their starting pool as the post-FYE cohort entering their program. This may be the best information available, but it invariably undercounts starters by failing to account for FYE students who leave the institution or switch to non-engineering majors. In the absence of the FYE requirement, some of these students would have been ME starters. By neglecting these students, the count of ME starters is artificially low resulting in an ME graduation rate that is artificially high. The same is true for every degree-granting engineering major in an FYE institution.

Because of the special nature of FYE programs, we cannot address starter miscounts by grouping FYE students with those admitted with "undecided" or "unknown" CIP codes—FYE students are neither. They were admitted as Engineering majors (2-digit CIP 14). However, we don't know to which degree-granting program (6-digit CIP) they intended to transition.

Therefore, to avoid miscounting starters at FYE institutions, we use "FYE proxies" which estimate the 6-digit CIP codes of the degree-granting engineering programs that FYE students would have declared had they not been required to enroll in FYE. We construct a data frame

suitable for imputation Multivariate Imputation by Chained Equations (MICE) algorithm using the mice R package [10]. For a given set of source files, FYE proxies need be created only once and written to file. The result can be used as needed unless the source files change.

See [11] for the R code that we use to determine FYE proxies and [12] for the R code that we use to determine starters in each major.

Step Six: Identify students who graduate

An undergraduate student who completes their program and earns their first bachelor's degree is a *completer*. To be counted among their program's *graduates* however usually depends on whether they satisfy the criterion for *timely completion*. We derive a *completion status* variable to filter student-level records to obtain a bloc of graduates.

The next step might be to subset the graduates if necessary to meet the needs of the metric. For example, the graduation rate metric requires graduates to be a subset of starters in the same program. We postpone this step until describing the metrics later in the paper.

See [13] for the R code that we use to identify students who graduate.

Step Seven: Create grouping variables

We add grouping variables from the MIDFIELD data tables to our blocs in progress. We select these variables to provide the aggregating categories we want for a particular metric. Program labels and student demographics are two of the most common sets of grouping variables we use. See [14] for the R code that we use to create groupings.

Program labels

At this point in a typical workflow, we have a bloc of student-level records in progress and a data frame of program labels. Both data frames have a 6-digit CIP variable to join by.

Program labels serve two main functions:

- *Filtering variable* to finalize a bloc. For example, "starters" or "graduates" usually mean starters or graduates *in specific programs*. Thus a bloc procedure typically concludes with a program filter as in ever-enrolled, starters, or graduates.
- *Grouping variable* for summarizing data. Having filtered a bloc to retain records in specific programs, the program label is retained and used with other grouping variables such as race/ethnicity and sex when computing and comparing metrics.

Rationale for the inner join. An inner join accomplishes two tasks: adds a column of program labels to the bloc; and filters the bloc to retain only those observations with CIPs matching the desired programs.

Demographics

Demographic variables (race/ethnicity and sex) are regularly left-joined to blocs for grouping and summarizing. We often want to remove records for which race/ethnicity or sex are "unknown".

Add origin

Origin is a demographic variable we use to distinguish "domestic" students from "international" students. The variable is a recoding of the race variable since international students typically must choose "international" rather than another race category at most institutions.

Add people

People is a demographic variable we use in many of our summaries. The variable combines the race and sex variables.

Add people by origin

Combining the two ideas above, again assuming that the observations on unknown race/ethnicity and sex have been removed,

Add Other variables

Depending on one's research question, any number of MIDFIELD variables might be used for grouping records.

Analysis metric 1: Graduation rate

Graduation rate—the fraction of a cohort of program starters who complete their program in a timely manner (typically 6 years)—is a widely used, though flawed, measure of academic achievement. The American Council on Education estimates that the conventional definition of graduation rate may exclude up to 60% of students at 4-year institutions. Nevertheless, as Cook and Hartle [15] explain,

> … in the eyes of the public, policy makers, and the media, graduation rate is a clear, simple, and logical—if often misleading—number.

Recognizing that graduation rate is a popular metric, we propose a definition of graduation rate that includes all conventionally excluded students except migrators.

Starters and migrators

As they pertain to the graduation rate metric, relationships among starters, migrators, and graduates (timely completers) of a given program *P* are illustrated in Figure 3.

- The overall rectangle represents the set of students ever enrolled in program *P*.
- The interior rectangle represents the set of graduates (timely completers) of program *P*.
- Region 1 (shaded) represents the graduation rate denominator, the set of starters in program *P*.
- Region 2 (shaded) represents the graduation rate numerator, the subset of starters who are also graduates of program *P*.
- Region 3 (unshaded) represents the set of students excluded from the graduation rate metric, depending on how "program" is defined as discussed below.

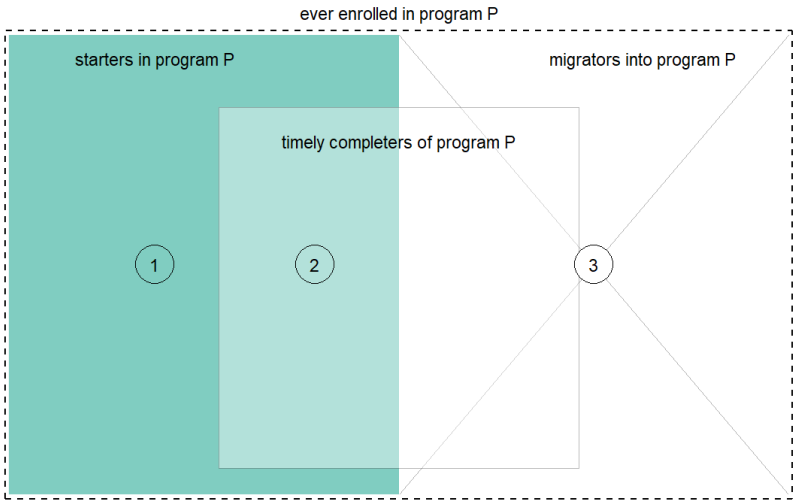

Figure 3. Graduation rate metric. Starters, migrators, and timely completers.

When calculating graduation rate, whether migrator-graduates are included in the count of graduates depends how a program is defined in terms of CIP codes.

- *Institution level*. Graduation rate computed at the institution level includes all migrators within the institution. For example, starters in Engineering (CIP 14) who graduate in Business (CIP 52) are both starters and timely completers at the institution level. IPEDS defines this rate as the *institution completion rate.*
- *2-digit CIP*. Graduation rate includes migrator graduates within the same 2-digit CIP. For example, starters in Engineering (CIP 14) graduating in Business (CIP 52) are excluded from the count of Business graduates, but migrators within Engineering (all 6-digit CIP codes starting with 14) are both starters and timely completers in Engineering.
- *4-digit CIP*. Similar to the 2-digit case. For example, starters in Electrical Engineering (CIP 1410) graduating in Mechanical Engineering (CIP 1419) are excluded from the count of Mechanical Engineering graduates, but migrators within Electrical Engineering (all 6-digit CIP codes starting with 1410) are both starters and timely completers in Electrical Engineering.
- *6-digit CIP*. Rarely used. Graduation rate at this CIP level excludes all migrators from the count of graduates.

- *Multiple CIPs*. In some cases, a single program or major includes different 4-digit CIPs. For example, migrators between Systems Engineering (CIP 1427), Industrial Engineering (CIP 1435), Manufacturing Engineering (CIP 1436), and Operations Research (CIP 1437) might be considered both starters and timely completers in a general program of Industrial & Systems Engineering.

Who is a starter?

In the US, the predominant definition of graduation rate is that established by the US Department of Education, Integrated Postsecondary Education Data System (IPEDS). The IPEDS definition underlies the finding cited earlier that a graduation rate metric may exclude up to 60% of students.

Many of the IPEDS exclusions relate to how starters are defined. By expanding the starters definition, MIDFIELD proposes a graduation rate definition that includes all conventionally excluded students except migrators.

Table 1: Comparing graduation rate definitions

Item	IPEDS	MIDFIELD	MIDFIELD notes
completion span:	4, 6, or 8 years	4, 6, or 8 years	Typical usage is 6 years
students admitted in:	Summer/Fall only	any term	
part-time students are:	excluded	included	Timely completion same as full-time students
transfer students are:	excluded	included	Timely completion span adjusted for level at entry

See [16] for the R code that we use to create graduation rate.

Analysis metric: Stickiness

Stickiness is a more-inclusive alternative to graduation rate as a measure of a program's success in attracting, keeping, and graduating their undergraduates. Stickiness is the ratio of the number of graduates of a program to the number ever enrolled in the program. All students excluded by a conventional graduation rate metric–including migrators—are included in the stickiness metric [17]. See [18] for the R code that we use to create stickiness.

Stickiness, in comparison to graduation rate, has these characteristics:

- Includes migrators, where graduation rate does not.
- Is based on the bloc of ever enrolled rather than starters, so there is no need for FYE proxies.
- Counts all graduates (timely completers) in a program, eliminating the need to filter graduates based on their starting program.

- Like the MIDFIELD definition of graduation rate (in contrast to the IPEDS definition), includes students who attend college part-time, who transfer between institutions, and who start in any term.

As they pertain to the stickiness metric, relationships among starters, migrators, and graduates (timely completers) of a given program *P* are illustrated in Figure 4.

- The interior rectangle represents the stickiness numerator, the set of graduates (timely completers) of program *P*.
- The overall rectangle represents the stickiness denominator, the set of students ever enrolled in program *P*.

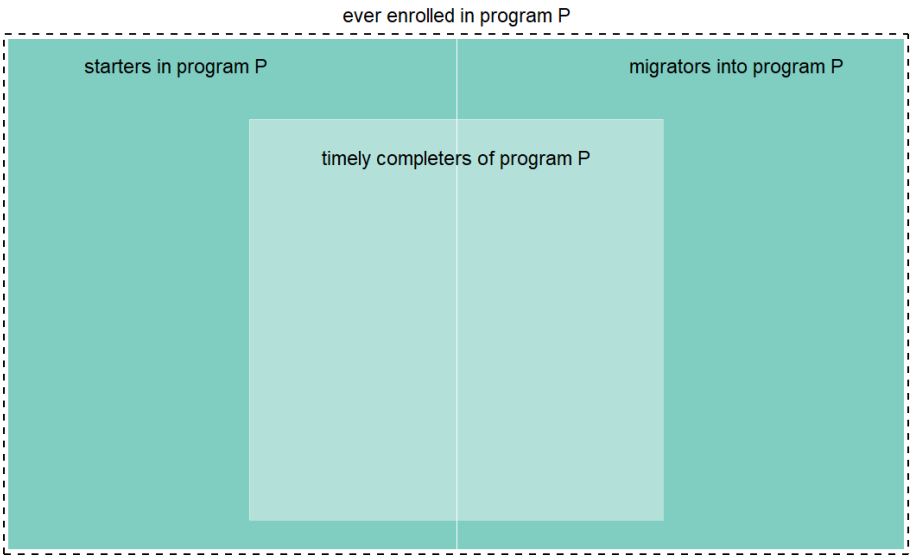

Figure 4. Stickiness metric. Starters, migrators, and timely completers.

Discussion and Conclusion

This paper has introduced the landscape of multi-institution longitudinal research and has provided tips for working with this valuable resource. This is a methodology that has been developed over time and is our best explanation of the process we use to prepare the dataset for analysis. We have described our method in detail and used the examples of graduation rate and stickiness as metrics for analysis. Our method is a sequence – however the steps do not need to be taken in the order that we have described here. Though our research uses MIDFIELD, the steps demonstrated can be used when preparing any longitudinal student database for analysis.

We hope this primer is helpful for the research community. Our research using these approaches has enabled impactful and award-winning research and informed institutional decision-making. For examples, see [2, 3, 4]. We encourage institutions to make evidence-based decisions rather than relying on anecdotal information. We recognize using robust methods for analyzing data can be challenging but the challenges are worth the effort to enact real change.

Acknowledgements

This work is generously supported through NSF Award Numbers 2142087 "Collaborative Research: Sustaining and Scaling the impact of the MIDFIELD project at the American Society for Engineering Education" and NSF Award 1545667 "Expanding Access to and Participation in the Multiple Institution Database for Investigating Engineering Longitudinal Development."

References

[1] "MIDFIELD | Multiple Institution Database for Investigating Engineering Longitudinal Development." https://midfield.online (accessed April 24, 2023).

[2] MIDFIELD. "Publications – MIDFIELD." https://midfield.online/publications/ (accessed April 24, 2023).

[3] M. W. Ohland and R. A. Long, "The Multiple-Institution Database for Investigating Engineering Longitudinal Development: An Experiential Case Study of Data Sharing and Reuse," *Advances in Engineering Education,* vol. 5 no. 2, pp. 1-25, 2016. https://advances.asee.org/wp-content/uploads/vol05/issue02/Papers/AEE-18-Ohland.pdf

[4] S. M. Lord, M. W. Ohland, M. K. Orr, R. A. Layton, R. A. Long, C. E. Brawner, H. Ebrahiminejad, B. A. Martin, G. D. Ricco, and L. Zahedi, "MIDFIELD: A Resource for Longitudinal Student Record Research," *IEEE Transactions on Education*, vol. 65, no. 3, 245-256, 2022. DOI 10.1109/TE.2021.3137086

[5] National Center for Education Statistics. "CIP user site." https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=55 (accessed April 24, 2023).

[6] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Data sufficiency – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-020-data-sufficiency.html (accessed April 24, 2023).

[7] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Degree seeking – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-030-degree-seeking.html (accessed April 24, 2023.

[8] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Programs – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-040-programs.html (accessed April 24, 2023).

[9] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Blocs – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-050-blocs.html (accessed on April 24, 2023).

[10] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations," *R. Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011 . https://doi.org/10.18637/jss.v045.i03

[11] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "FYE proxies." https://midfieldr.github.io/midfieldr/articles/art-060-fye-proxies.html (accessed April 24, 2023).
[12] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Starters – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-070-starters.html (accessed April 24, 2023).
[13] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Graduates – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-080-graduates.html (accessed April 24, 2023).
[14] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Groupings – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-090-groupings.html (accessed April 24, 2023).
[15] B. Cook and T.W. Hartle. "Why graduation rates matter – and why they don't." *The Presidency Magazine*, vol. 14, no. 2, pp. 32-35, 2011.
[16] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Graduation rate – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-100-grad-rate.html (accessed April 24, 2023).
[17] M. Ohland, M. Orr, R. Layton, S. Lord, and R. Long. "Introducing stickiness as a versatile metric of engineering persistence." In *Proceedings of the 2012 Frontiers in Education Conference*, 1–5. DOI 10.1109/FIE.2012.6462214.
[18] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord. "Stickiness – midfieldr." https://midfieldr.github.io/midfieldr/articles/art-110-stickiness.html (accessed April 24, 2023).

Appendix 1: Glossary

bloc	A grouping of student-level data dealt with as a unit, for example, starters, students ever-enrolled, graduates, transfer students, traditional and non-traditional students, migrators, etc.
CIP	*Classification of Instructional Programs*, a taxonomy of academic programs curated by the US Department of Education [5]. The 2010 codes are included with midfieldr in the data set cip.
cip6	Character variable in the term and degree data tables of program observations. Values are 6-digit CIP codes.
completers	Bloc of students who complete their baccalaureate programs, earning their first degrees.
completion status	A derived midfieldr variable indicating whether a student completes a degree, and if so, whether their completion was timely. Possible values are "timely", "late", and "NA". Late completers are often excluded from a count of "graduates."
data range	The overall span of academic terms of student unit record data provided by an institution. We are particularly interested in the lower and upper limits of a continuous range.
data sufficiency criterion	Student records are limited to those for which available data are sufficient to assess timely completion without biased counts of completers or non-completers.
entry term	A student's first term in the database.
ever-enrolled	Bloc of students whose term records include a specified program in at least one term.
FYE	First-Year Engineering program, a common-first-year curriculum that is a prerequisite for declaring an engineering major at some US institutions. Denoted by its own CIP code (14.0102), FYE is not a degree-granting program.
FYE proxy	Our estimate of the degree-granting engineering program in which an FYE student would have enrolled had they not been required to enroll in FYE. The proxy, a 6-digit CIP code, denotes the program of which the FYE student can be considered a starter.
graduates	Bloc of all graduates (timely completers) from a program, without regard to their starting programs.
graduation rate	Graduation rate (G) is the ratio of the number of program "starter-graduates" (Nsg) (i.e., graduates from the program in which they started) to the number of program starters (Ns). G= Nsg/ Ns
graduation rate (IPEDS)	The fraction of a cohort of full-time, first-time, degree-seeking undergraduates who complete their program within a percentage (100%, 150%, or 200%) of the "normal" time (typically 4 years) as defined by the institution. IPEDS excludes students who attend college part-time, who transfer between institutions, and who start in Winter or Spring terms.

graduation rate (MIDFIELD)	The fraction of a cohort of degree-seeking undergraduates who complete their program in a timely manner (typically 6 years). MIDFIELD includes students who attend college part-time, who transfer between institutions, and who start in any term.
migrators	Bloc of students who leave one program to enroll in another. Also called *switchers.*
multiple imputation	Method of imputing missing categorical data, in this case, imputing the FYE proxy 6-digit CIP codes.
program	US academic field of study. Can be used to indicate a specialty within a field or a collection of fields within a Department, College, or University. Programs are denoted by the *Classification of Instructional Programs* (CIP), a taxonomy of academic programs curated by the US Department of Education [5].
start term	The first term in which a student can be considered a starter. Identical to the entry term unless the student enters as undecided/unspecified.
starters	Bloc of degree-seeking students in their initial terms enrolled in degree-granting programs.
starter-graduates	Subset of the starters bloc who are graduates (timely completers) from their starting programs.
stickiness	Stickiness is the ratio of the number of graduates of a program to the number ever enrolled in the program.
student-level data	Data at the "student-level" refers to information about individual students including, for example, demographics, programs, academic standing, courses, grades, and degrees.
timely completion criterion	Completing a program in no more than a specified span of years, in many cases, within 6 years after admission (150% of the "normal" 4-year span), or possibly less for some transfer students.
timely completion term	The last term in which a student's degree completion would be considered timely. In many cases the timely completion (TC) term is 6 years after admission. The TC term can be adjusted to account for transfer credits. (Currently, there is no mechanism for extending the TC term for co-ops or migrators.)
undecided/unspecified	The MIDFIELD taxonomy includes the non-IPEDS code (CIP 999999) for Undecided or Unspecified indicating instances in which a student has not declared a major or an institution had not recorded a program.