BOARD # 328: BPE: Year Three of Developing a New Dataset for Analyzing Engineering Curricula

Dr. David Reeping, University of Cincinnati

Dr. David Reeping is an Assistant Professor in the Department of Engineering and Computing Education at the University of Cincinnati. He earned his Ph.D. in Engineering Education from Virginia Tech and was a National Science Foundation Graduate Research Fellow. He received his B.S. in Engineering Education with a Mathematics minor from Ohio Northern University. His main research interests include transfer student information asymmetries, threshold concepts, curricular complexity, and advancing quantitative and fully integrated mixed methods.

Dr. Matthew W. Ohland, Purdue University at West Lafayette (PWL) (COE)

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 and represents ASEE on the Engineering Accreditation Commission. He was the 2002–2006 President of Tau Beta Pi and is a Fellow of the ASEE, IEEE, and AAAS. He was inducted into the ASEE Hall of Fame in 2023.

Dr. Kenneth Reid, University of Indianapolis

Kenneth Reid is the Associate Dean and Director of Engineering at the R. B. Annis School of Engineering at the University of Indianapolis. He and his coauthors were awarded the Wickenden award (Journal of Engineering Education, 2014) and Best Paper award, Educational Research and Methods Division (ASEE, 2014). He was awarded an IEEE-USA Professional Achievement Award (2013) for designing the B.S. degree in Engineering Education. He is a co-PI on the ”Engineering for Us All” (e4usa) project to develop a high school engineering course ”for all”. He is active in engineering within K-12, (Technology Student Association Board of Directors) and has written multiple texts in Engineering, Mathematics and Digital Electronics. He earned a PhD in Engineering Education from Purdue University, is a Senior Member of IEEE, on the Board of Governors of the IEEE Education Society, and a Member of Tau Beta Pi.

Dr. Hossein EbrahimNejad, Drexel University

Hossein EbrahimiNejad is a data scientist currently working with the office of Enrollment Analytics at Drexel University. He received his PhD in Engineering Education from Purdue University, where he gained a strong knowledge of higher education and strategic enrollment management. Hossein's skills in data management, data visualization, and predictive modeling allow him to empower stakeholders to make strategic decisions using advanced analytic information. With Drexel's commitment to promoting and supporting student success, Hossein's work with Enrollment Analytics allows the University to make data-driven strategic decisions regarding enrollment and financial projections that are as effective and efficient as possible.

NAHAL RASHEDI, University of Cincinnati

PhD Student of Engineering Education

Yunmeng Han, University of Cincinnati

Yunmeng Han is a PhD student in Engineering Education at the University of Cincinnati. She holds a Master's degree in Computer Science from Northeastern University and serves as a reviewer for prominent engineering education conferences, including ASEE 2025. Yunmeng has been actively involved in NSF-funded research projects and is experienced in applying both quantitative and mixed-method research designs.

BPE: Year Three of Developing a New Dataset for Analyzing Engineering Curricula

Abstract

This paper presents progress from Year 3 of a National Science Foundation-supported Broadening Participation in Engineering project, which investigates structural barriers contributing to the attrition of diverse student subgroups in engineering. Specifically, we focus on curricular factors using an emerging network analysis framework that quantifies the "complexity" of engineering curricula. Our study leverages a dataset of 497 plans of study across five engineering disciplines (Mechanical, Civil, Electrical, Chemical, and Industrial) from 13 institutions represented in the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) – and the database itself. MIDFIELD includes course-taking records and demographic information for all students enrolled at these institutions. Year 3 of this project focused on our final proposed tasks: analyzing course-taking trajectories in the MIDFIELD data to find patterns across groups of students, finalizing a scoping review of additional metrics to facilitate curricular analyses, and distributing an R package to employ the metrics found in the scoping review and associated analyses conducted throughout this project.

Context of Project

This project integrates two approaches to studying student progression in engineering: the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [1] and a framework for assessing a curriculum's "complexity" known as Curricular Analytics [2]. MIDFIELD is a widely used resource in engineering education research, particularly for examining retention across disciplines. The data contains a wealth of information for each student, including demographics, academic standing, major, course records, GPA, and graduating term. Although MIDFIELD has been primarily studied in terms of retention and graduation, the student course records remain relatively underexplored, which offers a promising opportunity to synergize with an emerging framework for analyzing curricula, Curricular Analytics. Introduced in its current form by Heileman et al. [2], Curricular Analytics models a degree program's plan of study as a network, with courses represented as nodes and pre and corequisites as edges connecting them. This type of representation allows us to explicitly capture all the dependencies underlying a student's journey toward a degree in engineering using network analysis, which provides quantitative evidence suitable for comparison.

Curricular Analytics is most concerned with two metrics: (1) the *blocking factor*, which is the number of courses rendered inaccessible if a given course is failed and (2), the *delay factor*, which is the length of the longest prerequisite chain that includes a given course. When these metrics are added together, we obtain what's called the *cruciality.* The cruciality can be interpreted as a local measure of the course's importance in a particular plan of study. To get a global measure of the curriculum's complexity, each of the crucialities can be summed to yield the *structural complexity.* Structural complexity has been linked to program completion rates—higher structural complexity is associated with lower completion rates [2], [3], [4].
Within Curricular Analytics, the latent factors of a curriculum, such as the availability of peer tutoring programs, instructional quality, and academic advising, are also captured in what's called

instructional complexity. Despite being a vital component of what drives student flow through a curriculum, instructional complexity has not seen the level of attention theoretically or analytically as the structural complexity metric (c.f., [5]). The current proxy for instructional complexity is the pass rate for a course [2], which is the minimum data needed to model student flow using techniques like Markov Chains or Agent-Based Modeling (ABM). To be theoretically complete, more work is needed to operationalize instructional complexity.

Research Questions

In our project, we combine the course-taking data in MIDFIELD with the plan of study data for five disciplines of engineering (Mechanical, Electrical, Civil, Chemical, and Industrial) at 13 institutions [6]. We also included a longitudinal element in our data collection by extracting program data that looked back ten years from the most recent entry in MIDFIELD. Our primary research question in this project is: *How does the complexity of the codified curriculum vary among institutions and disciplines?* With student course-taking data, we are expanding upon this question by addressing the following questions: (1) *How do different populations and pathways (e.g., FTIC, changing majors) navigate the curriculum?;* (2) *To what extent do students follow the codified curriculum?;* and (3) *How is the curricular complexity experienced by students related to overall GPA, discipline stickiness, and migration yield?*

Summary of Year 2 Project Activities

In Year 2, we expanded on the work in Year 1, which was almost exclusively collecting data to build a plan of study dataset amenable to Curricular Analytics and developing an R package to facilitate customized analyses with the dataset. We focused on three primary tasks in Year 2: verification of the plan of study dataset and calculating the descriptive statistics for the data, expansion of the R package through a systematic literature review, and the analysis of curricular design patterns.

Verification of the Plan of Study Dataset and Descriptive Statistics. Much of our time in Year 2 was spent ensuring the accuracy of the data after running preliminary analyses and noting anomalies in the data. We used Python to check for inconsistencies in prerequisite structures, such as mismatched or missing prerequisites, across disciplines and years. These were then manually corrected by referring to the institutional catalog. The corrections increased the mean structural complexity scores by approximately 2% [7]. The mean structural complexity is 319, and the median is 301. The smallest structural complexity we observed was 122, whereas the largest was 897. Chemical engineering exhibited the largest mean structural complexity of 436, followed by mechanical engineering with 374. The means of the remaining disciplines – electrical, industrial, and civil – were much closer to one another (295, 257, and 240, respectively).

Expansion of the R Package. The R package developed in Year 1 to scale Curricular Analytics was planned to be enhanced with additional metrics identified through a scoping literature review (SLR). The SLR examined 159 papers and identified 61 that expanded on Curricular Analytics, after which metrics were extracted from the papers. At the time, the metrics were categorized by type (structural or instructional) and level (student, course, or curriculum).

Analysis of Curricular Design Patterns. Finally, once the data were verified, we explored curricular design patterns across the data by parsing the plans of study into components such as the Calculus sequence and first-year engineering programs (see [8]). Courses were standardized to allow for cross-institutional comparisons, leveraging GPT-4 to generalize the course names.

Major Activities During Year 3

Our Year 3 efforts have shifted to spotlight the possible insights to be gained from MIDFIELD in light of the new dataset we collected while closing the loose ends in Year 2. Accordingly, our activities centered on analyzing course-taking trajectories, finalizing our scoping review, and distributing the developed R package (in addition to the dataset and resources).

Analyzing Course-Taking Trajectories. At the onset of Year 3, we began work on the following research question: *How do different populations and pathways (e.g., FTIC, changing majors, transfer) navigate the curriculum?* To address this question, we planned to use association analysis to discover frequent groupings of courses and association rules among them to build course-taking trajectories. The concept of association analysis [9] is classically applied to analyzing transaction data to observe what items are bought together and develop *association rules* of the form, "the people who bought item(s), A, also tended to buy item(s), B." Those bundles of items we are trying to relate, A and B, are called *itemsets*, and the association rule would be written as $A \rightarrow B$ (A is the antecedent and B is the consequent). The output of these analyses looks like suggestions provided by online retailers like Amazon, where each product page is supplemented by a "Customers also bought these items" section. Instead, the output here is "Students also took these courses."

We intended to use association analysis to identify typical course-taking trajectories for various groups of students. However, the results of our initial application of the technique did not unearth deviant pathways to the extent we needed because of lower frequencies relative to the dominant path. Instead, we are leveraging an emerging clustering technique that can process data with a dimensionality that far exceeds the sample size, Thresholding After Random Projects (*n*-TARP) [10], [11]. This technique projects the data to a one-dimensional space by multiplying each data point by a random vector using a dot product; k-means is then run to form two clusters. Because the technique uses random vectors to generate the one-dimensional projection, n-TARP produces a collection of possible solutions. Although *n*-TARP only produces two clusters in a single solution, the researcher can rerun the analysis on specific clusters to uncover sub-patterns.

We have converted the instances of students' course-taking into a bag-of-words representation, a vector containing each instance of a course being taken by the student, which includes courses that were retaken. We are currently experimenting with clustering student trajectories at single institutions to iteratively split clusters of students who start in a specific year to identify trajectories to completing an engineering degree – or not. To explore differences between clusters, we are comparing the normalized frequency of the courses across the clusters, the number of major switchers, transfer status, and other demographic categories. We expect to have results of the *n*-TARP clustering by the time of the conference.

Finalizing Scoping Review. Starting in Year 2, we conducted a scoping review to identify additional metrics that researchers and practitioners could use with our dataset and their data [7]. The metrics were drawn from 61 papers citing foundational Curricular Analytics research since its introduction in 2013. Figure 1 provides a visual summary of our review.

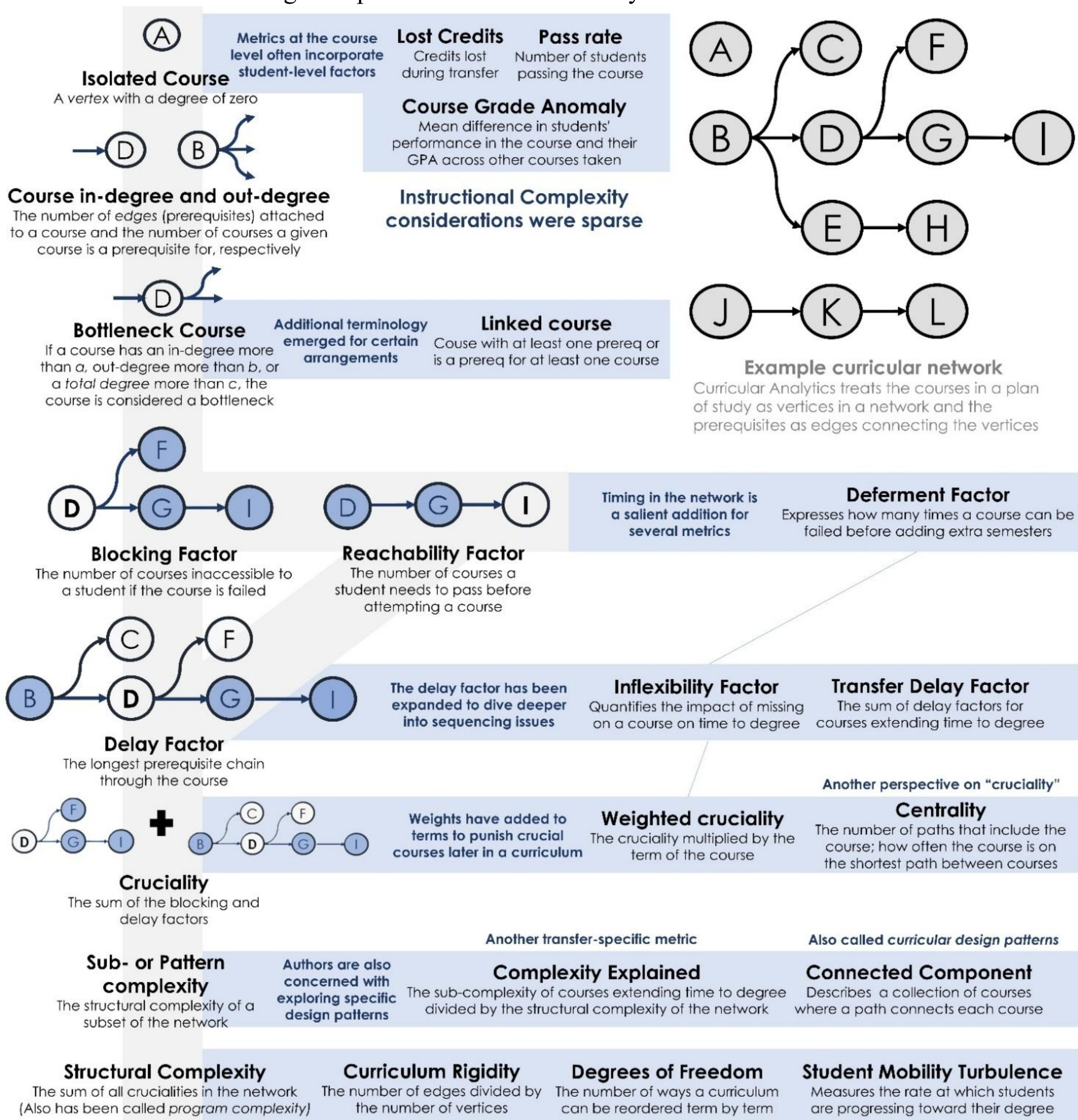

Figure 1. Summary of scoping review with selection of metrics

We extracted 47 unique metrics from the papers. While mapping the relationships between the metrics, we found there is a lack of instructional complexity metrics in the literature, and those that exist do not fully capture the intention of instructional complexity.

Developing and Distributing an R Package. Based on the scoping review results, we have added metrics like *bottleneck course, centrality, curriculum rigidity,* and *deferment factor* to our package and plan to add more. The package, including the dataset, will be available on GitHub by the closure of the project. In the meantime, we plan to host workshops on using the package through the network of MIDFIELD researchers and at a future ASEE and Frontiers in Education conference.

Future Work and Conclusion

Next, we plan to address the last two questions in our project: *To what extent do students follow the curriculum as codified?* And *how the curricular complexity experienced by students is related to overall GPA, discipline stickiness, and migration yield.* These will be addressed through correlational analyses. By deconstructing the varied pathways students take to an engineering degree, we can better understand what curricular bottlenecks exist for students and find appropriate ways to increase the flexibility of our programs.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. EEC-2152441. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] S. M. Lord *et al.*, "MIDFIELD: A Resource for Longitudinal Student Record Research," *IEEE Trans. Educ.*, vol. 65, no. 3, pp. 245–256, Aug. 2022, doi: 10.1109/TE.2021.3137086.

[2] G. L. Heileman, C. T. Abdallah, A. Slim, and M. Hickman, "Curricular Analytics: A Framework for Quantifying the Impact of Curricular Reforms and Pedagogical Innovations," *ArXiv181109676 Phys.*, Nov. 2018, Accessed: Aug. 04, 2021. [Online]. Available: http://arxiv.org/abs/1811.09676

[3] A. Slim, "Curricular Analytics in Higher Education," Dissertation, The University of New Mexico, 2016. Accessed: Feb. 24, 2023. [Online]. Available: https://www.proquest.com/docview/1873863748?pq-origsite=gscholar&fromopenview=true

[4] D. M. Grote, D. B. Knight, W. C. Lee, and B. A. Watford, "Navigating the curricular maze: Examining the complexities of articulated pathways for transfer students in engineering," *Community Coll. J. Res. Pract.*, vol. 45, no. 11, pp. 779–801, 2021.

[5] D. Waller, "Organizational factors and engineering student persistence," Dissertation, Purdue University, 2022. [Online]. Available: https://doi.org/10.25394/PGS.21606342.v1

[6] D. Reeping, S. M. Padhye, and N. Rashedi, "A Process for Systematically Collecting Plan of Study Data for Curricular Analytics," presented at the American Society for Engineering Education Annual Conference, Baltimore, MD, 2023. doi: https://peer.asee.org/42466.

[7] D. Reeping, K. Reid, M. Ohland, and N. Rashedi, "Board 438: Year Two of Developing a New Dataset for Analyzing Engineering Curricula," presented at the American Society for Engineering Education Annual Conference, Portland, OR, 2024.

[8] S. Padhye, D. Reeping, and N. Rashedi, "Analyzing Trends in Curricular Complexity and Extracting Common Curricular Design Patterns," presented at the American Society for Engineering Education Annual Conference, Portland, OR, 2024.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, in SIGMOD '93. New York, NY, USA: Association for Computing Machinery, Jun. 1993, pp. 207–216. doi: 10.1145/170035.170072.

[10] T. Yellamraju, "N-TARP: A Random Projection Based Method for Supervised and Unsupervised Machine Learning in High-dimensions with Application to Educational Data Analysis," *Theses Diss. Available ProQuest*, pp. 1–136, Jan. 2019.

[11] T. Yellamraju and M. Boutin, "Pattern Dependence Detection using n-TARP Clustering," Jun. 13, 2018, *arXiv*: arXiv:1806.05297. doi: 10.48550/arXiv.1806.05297.