Board 438: Year Two 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. 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. Matthew W. Ohland, Purdue University, West Lafayette

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.

NAHAL RASHEDI, University of Cincinnati

PhD Student of Engineering Education

Year Two of Developing a New Dataset for Analyzing Engineering Curricula

Abstract

This paper discusses the developments during Year 2 of a project concerned with analyzing the curricula of engineering programs in the United States to understand the structural barriers embedded in degree requirements that could push out diverse groups of students. We are using an emerging method for quantifying the complexity of these programs called Curricular Analytics. This method involves treating the prerequisite relationships between courses as a network and applying graph theoretic measures to calculate a curriculum's complexity. In Year 1, we collected 494 plans of study representing five engineering disciplines (i.e., Mechanical, Civil, Electrical, Chemical, and Industrial) across 13 institutions - spanning a decade. To ensure the dataset is as useful as possible to engineering education researchers, we have intentionally aligned our data collection with institutions available in the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD). One of the outputs of this project is an R package that will enable researchers and practitioners to explore and leverage the new dataset in their work by enabling the calculations to be completed at scale. With the efforts in Year 1, the package has the functionality to compute the necessary metrics for Curricular Analytics. We are currently conducting a systematic literature review of how Curricular Analytics has been applied and extended to search for other promising metrics to add to our package. This paper will provide an update on the preliminary analyses we have conducted using Curricular Analytics, an introduction to the R package, and updates from our systematic literature review.

Context of the Project

This project is a combination of two tools for understanding student progression in engineering: the Multiple Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [1] and a framework for quantifying the "complexity" of a curriculum called *Curricular Analytics* [2]. The MIDFIELD dataset is a popular resource for studying retention across disciplines in engineering education research. In particular, the dataset contains the following "tables": (1) a course table including each instance of a course attempted by a student; (2) a term table that captures the student's academic standing and the program in which they are enrolled; (3) a demographics table describing student's relevant background information; and (4) a degree table cataloging the graduating students' term and the program they completed [3]. Of these data, however, the course table is relatively untouched. This collection of untapped data is where the opportunity for the second tool, Curricular Analytics, enters. Introduced formally by Heileman et al. [2] in its current iteration, the framework treats a plan of study for a degree program as a network, where each course is represented as a node and pre- and corequisites are given by edges that connect courses together. By representing curricula in a network data structure, it is able to be explored using a suite of network analysis techniques.

Curricular analytics is most concerned with two metrics: the blocking factor and the delay factor. For each course, the blocking factor is a count of the number of courses that are inaccessible to the student if the course in question is failed. The delay factor concerns the length of the prerequisite chains flowing through the course; the length of the longest prerequisite chain that includes the course in question is the value of the delay factor. Adding the delay factor and blocking factor together yields what is called the cruciality. The cruciality provides a numerical indicator of how essential the course is to the plan of study with respect to completing program requirements [4].

Further summing the crucialities in a curriculum yields what is called *structural complexity.* This measure has gained attention for its relationship with completion rates; as structural complexity increases, completion rates decrease [2], [5], [6]. Another quantity, called instructional complexity, attempts to capture the latent factors in the curriculum, such as instructional quality, availability of peer tutoring, and other environmental factors contributing to student success. However, it has received much less attention because of the vast amount of data necessary to model successfully compared to the structural complexity, which only requires the plan of study and the requisite information to calculate. Moreover, little theoretical effort has been placed into scoping what data to collect, settling for pass rates of individual courses as the single variable to collect.

Given the robustness of the structural complexity measure in Curricular Analytics, we focus on examining the complexity of engineering curricula structurally. In fact, by combining curricular data with actual student course-taking trajectories as offered in MIDFIELD's course table, there is a considerable opportunity to unpack student behaviors with respect to their course-taking and inform future research on how the curriculum itself can be a barrier to student success.

Summary of Relevant Major Activities During Year 1

Our previous annual update [7] described how the first year was almost exclusively focused on data collection and development. MIDFIELD is an existing dataset that is ready to be analyzed, but no complementary dataset comprised of the curricular information necessary to apply Curricular Analytics existed at the time. Thus, we collected the plans of study for five engineering majors (i.e., Mechanical, Civil, Electrical, Chemical, and Industrial) starting at their most recent entry in MIDFIELD and tracing back nine years – capturing a decade of curricular information for each discipline. The data collection process is detailed in [7], wherein a team of five undergraduate research assistants and one PhD student entered the relevant curricular data for 13 institutions in the MIDFIELD dataset. Preliminary analyses were conducted to explore the new data for correctness by sampling extreme cases (i.e., outliers) using boxplots. After exploring the outliers, such as those with comparatively low complexity scores of less than 100, we found hidden special characters that resulted from encoding errors after copying and pasting information from web pages. These special characters were interfering with our ability to create the plan of study networks and perform the network analysis, so we needed to spend additional time making corrections during Year 2.

At the same time the data was being collected, we developed a package for the statistical programming environment, R, to scale the analysis of the curricular data and enable us to integrate more flexible analyses into our research design. After successfully replicating previous work in [8] using the package and unit testing the functions in the package, it was deemed to be appropriately validated for application in this project.

Major Activities During Year 2

The two activities in Year 1 directly led us to pursue three main strands of work in Year 2: verification of the plan of study dataset, expanding the R package through a systematic literature review, and mining curricular design patterns.

Verification of Plan of Study Dataset. To ensure the dataset was as accurate as possible in its reflection of the curricular realities when the plan of study was created, we performed verification using Python to examine the internal consistency of prerequisite structures in two ways. This process lasted several months, from mid-summer 2023 to the end of Fall 2023.

First, we explored whether courses with the same name in the same year were recorded as having the same prerequisites across different disciplines. In some cases, prerequisites could be recorded differently depending upon which page was being examined during data collection – e.g., the plan of study web page had one set of prerequisites that conflicted with the university catalog. Because different disciplines share a subset of courses, we took advantage of this fact to find instances where disagreement emerged during data entry. In cases of divergence in which prerequisites were recorded, the original plan of study documentation was revisited to resolve the discrepancy.

Another type of error we explored was whether a course was listed as a prerequisite or corequisite but was not listed in the plan of study. Again, we revisited the plan of study to ensure the correct prerequisite relationship was recorded in the data. Unlike the consensus-based verification, it was possible for an "error" not to be caused by an oversight during data entry because the course did indeed not appear in the plan of study. These false positives were often found with courses in the first year, such as Calculus 1, which technically have prerequisites but students are intended to be placed in these through mechanisms like a math placement test.

Once we completed the verification process, we analyzed the plan of study data (n = 494) with the conventional Curricular Analytics framework, primarily using descriptive statistics. This involved calculating the structural complexity of each plan of study and disaggregating them by discipline and institution. The average structural complexity was found to be 318, and the median was 300; these figures provide some of the first benchmarks for how engineering programs (i.e., Civil, Chemical, Electrical, Industrial, and Mechanical) compare in terms of complexity to one another. Chemical engineering has the highest mean structural complexity of 436, followed by mechanical engineering with a structural complexity of 374. The remaining disciplines were more tightly clustered together: electrical with 293, industrial with 257, and civil with 242. Our verification process resulted in our initial estimates of structural complexity changing slightly upward by approximately 2%; this did not impact any previous inferences.

Prior to this effort, it was not clear what the typical value of structural complexity would be for a given program. With this new dataset, we now better understand what this *typical* value may be for engineering programs.

Expanding the R Package through a Systematic Literature Review. In our first year, we developed a package for the statistical programming environment, R, to conduct Curricular Analytics at scale. During our second year, after discussing the application of the package with researchers in engineering education and other disciplinary areas, we found it prudent to explore additional metrics that could be used to analyze the dataset. Although it was not our original intention to add more metrics beyond the essential few in Curricular Analytics, this new direction was pursued to bolster the usability of the R package developed in the project's first year.

As a pragmatic study within the context of the larger project, we are conducting a systematic literature review (SLR) on how researchers have attempted to analyze a curriculum quantitatively using Curricular Analytics. We posed the following research questions: "What metrics do researchers use to quantify curricula that build off the premise of Curricular Analytics?" and "What kind of methodologies are used in conjunction with their metrics?" In other words, we were interested in finding what other network-based measures were being used to explore the complexity of engineering programs. We are also interested in how the framework has been extended, such as how the measures were being used in a broader methodology.

The first step of collecting data in this SLR involved determining the inclusion criteria [9]. We were most interested in papers that analyzed the curriculum as a whole using quantitative methods. Because of the nature of our data, we focused on applications that expand on the ideas presented in Curricular Analytics. Moreover, considering our study's context is in engineering, we also focused on similar applications in engineering programs.

Thus, our inclusion criteria were as follows:
• Must use quantitative or mixed methods
• The curriculum is the unit of analysis
• Must use or build upon the curricular analytics framework (e.g., network analysis)
• Must be written in English
• Context of the study includes engineering

To form our sample, we identified papers that cited the foundational papers on Curricular Analytics using Google Scholar's citation tracking feature. These are detailed in Table 1.
A total of 307 papers were extracted on August 30th, 2023. After accounting for duplicates, 159 papers remained.

Table 1. Source papers used in the systematic literature review

Source Author	Paper	Citations
Gregory L. Heileman	Restructuring Curricular Patterns Using Bayesian Networks.	1
Gregory L. Heileman	Efficient curricula: The complexity of degree plans and their relation to degree completion	6
Gregory L. Heileman	Crucial based curriculum balancing: A new model for curriculum balancing	8
Gregory L. Heileman	Guiding early and often: Using curricular and learning analytics to shape teaching, learning, and student success in gateway courses	12
Gregory L. Heileman	Does Curricular Complexity Imply Program Quality?	18
Gregory L. Heileman	Characterizing the complexity of curricular patterns in engineering programs	32
Gregory L. Heileman	The complexity of university curricula according to course cruciality	34
Gregory L. Heileman	Curricular analytics: A framework for quantifying the impact of curricular reforms and pedagogical innovations	37
Gregory L. Heileman	Curricular efficiency: What role does it play in student success?	38
Gregory L. Heileman	Employing Markov networks on curriculum graphs to predict student performance	43
Ahmad Slim	Curricular analytics in higher education	23
Ahmad Slim	Network analysis of university courses	33
Ahmad Slim	The Impact of Course Enrollment Sequences on Student Success	22

We are currently reviewing the full texts of the 159 papers and have processed 122 of them so far. Of the 122 papers, 61 papers met the inclusion criteria. We are extracting the research questions, methods employed, and metrics introduced as part of the analysis from each of the papers. Much like how duplicate papers appeared during the search process, some metrics appeared multiple times – occasionally with different names. For example, the blocking factor and delay factor appear several times because they are core features of Curricular Analytics.

We are categorizing these metrics at two levels: (1) whether the metric is related to instructional or structural complexity and (2) whether the metric is at the student level, course level, or curriculum level. For example, one structural complexity factor, curriculum rigidity, refers to the number of prerequisites divided by the number of courses in the plan of study [10], [11]. Curriculum rigidity is a curriculum-level measure that attempts to directly quantify the level of connectivity in the requisite structures, where a value less than 1 would indicate a simpler structure and a value greater than 1 implies the network is more intricate. In fact, those familiar with graph theory will recognize this as the beta index, just applied to the context of a curriculum. On the other hand, an instructional complexity factor at the course level would be course grade anomaly, which is the mean difference between the overall GPA of a student and the students' grades in the course of interest [12].

We plan to integrate this package with the existing "midfieldr" [13] and “midfielddata” [3] packages, which empower researchers with more streamlined functions to analyze student course-taking trajectories in MIDFIELD. We will integrate appropriate additional metrics and

functionalities based on the SLR we are conducting. These new metrics will further enrich the analytical capabilities available to researchers, enabling them to gain deeper insights into various aspects of curricular complexity and engineering student experiences in higher education. Currently, the package contains the following functions and draws from the igraph package [14] to handle network-based calculations (Table 2).

Table 2. Current functionality in R package implementing Curricular Analytics

Function Name	Functionality
admissibility_test	Automatically checks for data entry issues that would impact the Curricular Analytics measures.
blocking_factor	Calculates the number of courses inaccessible to a student if a course is failed.
create_plan_of_study	Accepts the curricular data imported from a .csv as a dataframe, then converts it into an igraph object.
delay_factor	Calculates the longest prerequisite chain through a given course.
find_inbound_courses	Finds all courses connected to a given course that are direct or indirect pre- or corequisites.
Find_outbound_courses	Finds all courses that have the given course as a direct or indirect pre- or corequisite.
plot_plan_of_study	Plots the plan of study (as an igraph object) with the courses ordered by term.
structural_complexity	Calculates the overall structural complexity of a plan of study and outputs a table of the blocking factors, delay factors, and crucialities for each of the courses.
subcomplexity_graph	Creates a subgraph based on a user-selected course. The subgraph will contain any courses connected to the specified course, directly or indirectly.
what_if	Calculates the result of deleting or adding a user-defined course, prerequisite, or corequisite.

Mining Curricular Design Patterns. Once we verified the dataset, we began parsing the plan of study data into curricular design patterns [15]. These curricular design patterns concern the arrangements of courses, such as the Calculus sequence or a First-Year Engineering Program. Just like we can calculate the structural complexity of the entire curriculum, the structural complexity of curricular design patterns can be calculated as well (by using the subcomplexity_graph and structural_complexity functions in our package); the negative relationship between complexity and completion rates still holds. By isolating these curricular design patterns and treating them like atoms that make up larger molecules – i.e., the curriculum – we can better understand how to reduce unnecessary complexities for future engineering students.

To mine the design patterns, we labeled courses with a generalized category to better fetch courses from the plan of study dataset. This process was necessary because similar courses like Calculus I are not called by the same name everywhere (e.g., Calculus I, Calculus of a Single Variable, and First Year Calculus). We employed the large language model, GPT-4, through its

API to standardize the names of the courses [16]. As with using any large language model, hallucinations are possible; therefore, the output was manually checked and corrected as needed.

The curriculum design patterns were extracted iteratively by selecting a course as a focal point and finding its associated courses through prerequisite or corequisite relationships. From there, these curriculum design patterns were abstracted to general course numbers (i.e., 1,2,3…) and clustered by type of courses (e.g., Calculus, Statics, Introduction to Engineering). The results of this effort are described in Padhye et al. [17]. Moreover, the functionality to extract sequencing has been integrated into the R package so that users can query specific courses and calculate quantities of interest about the sub-network of courses. To illustrate how different curricular design patterns emerge, consider the two following examples focused on the course, Statics, a common bottleneck in engineering curricula [18].

In Figure 1, the curricular design pattern on top has two direct prerequisites, three indirect prerequisites, and one corequisite. Moreover, Statics has 12 courses that are blocked if it is failed. However, the curricular design pattern on the bottom presents a more direct path, with only 1 prerequisite. Although nearly as many courses are blocked (i.e., 11 courses), the overall structural complexity represents a difference of 79 points (242 versus 163, respectively) – this represents a 48% increase in complexity from the bottom curricular design pattern to the top. These are the kinds of data we can leverage to understand bottlenecks in the curricula with a finer-grained analytical method like Curricular Analytics. Our Year 3 activities will be focused on filtering through the curricular design patterns.

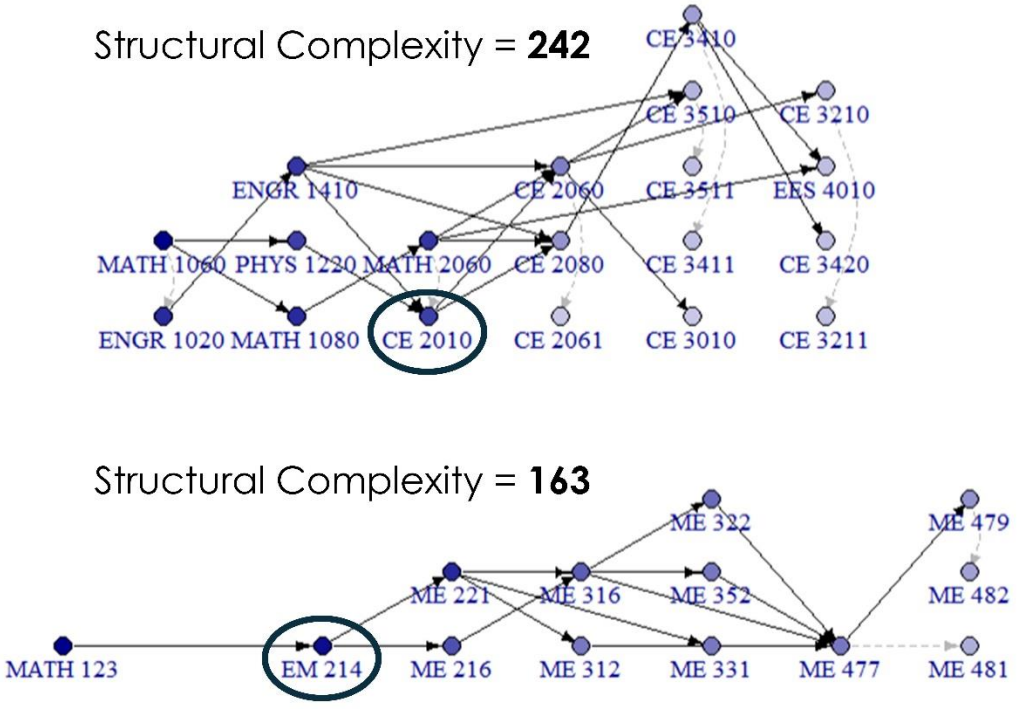

Figure 1. Two curricular design patterns focused on Statics (circled)

Planned Next Steps

Beyond completing the analyses described for our Year 2 activities, there are two primary activities that we expect to take place in Year 3: using association analysis to extract student course-taking trajectories to compare to the codified curricula and expanding access to the dataset and package.

Association Analysis of Student Course-Taking Trajectories. For Year 3, we plan on building functions to manipulate course-taking trajectories of actual student data so that they can be compared using association analysis. Association analysis is a data mining technique that enables a researcher to extract patterns of items, like courses, that are frequently taken together or in sequence. The classic example of association analysis is based on transactions in a grocery store, where the problem is to determine which items customers frequently purchase together. Thus, we can use the technique to build common course-taking trajectories for different groups of students. Association analysis will enable us to mine common course-taking patterns disaggregated by strata like institution, discipline, first-generation-status, and transfer-status and reconstruct them as networks to complement the plan of study data. This is the last step in our research design to complete the integration of the new curricular data with the student course-taking data in MIDFIELD.

Expanding Access to the Dataset and Package. We plan to engage the broader community in exploring the new dataset and package we created as part of this effort. At a future American Society for Engineering Education, American Education Research Association, and Frontiers in Education conference, we will host workshops to engage researchers across disciplines with these new tools. Moreover, we will provide an overview of how to intersect these data with MIDFIELD. Through interactions with others who share research interests within this area, we anticipate that collaborations can be formed to interrogate the data from different perspectives and increase the diversity of the institutions and disciplines in the sample.

We have begun distributing the package by request to researchers and will be making the dataset available as well. To increase accessibility to the dataset and package, an R Shiny application will be created with appropriate documentation to guide users on the intended uses of the tools and limitations to note when working with them (e.g., five engineering disciplines represented that are not offered at each institution).

Conclusion

Comprehending the intricacies of the curriculum can aid in optimizing programs for students and ensuring their timely graduation. By scrutinizing the influence of the curriculum on student progress with Curricular Analytics, supplemented by the dataset we created, researchers can uncover the barriers students encounter as they advance through their studies. These insights enable the identification of unnecessary complexities within the curriculum or areas where students may veer off course. Interventions based on the results might entail adjustments to curricular guidelines, enhanced academic advising, or the implementation of novel programs and initiatives to bolster student progress. As this project continues to evolve, we expect to deliver

new analytical potential to the community and create new strands of inquiry to connect to existing persistent problems in engineering education.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. BPE- 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] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord, "midfielddata: MIDFIELD data sample." 2022. [Online]. Available: https://midfieldr.github.io/midfielddata/

[4] A. Slim, J. Kozlick, G. L. Heileman, and C. T. Abdallah, "The Complexity of University Curricula According to Course Cruciality," in *2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems*, Birmingham, UK: IEEE, Jul. 2014, pp. 242–248. doi: 10.1109/CISIS.2014.34.

[5] 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

[6] 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.*, pp. 1–30, Aug. 2020, doi: 10.1080/10668926.2020.1798303.

[7] D. Reeping, S. M. Padhye, and N. Rashedi, "A Process for Systematically Collecting Plan of Study Data for Curricular Analytics," in *2023 ASEE Annual Conference & Exposition*, 2023.

[8] D. Reeping, D. M. Grote, and D. B. Knight, "Effects of large-scale programmatic change on electrical and computer engineering transfer student pathways," *IEEE Trans. Educ.*, vol. 64, no. 2, pp. 117–123, 2020.

[9] M. Borrego, M. J. Foster, and J. E. Froyd, "Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields: Systematic Literature Reviews in Engineering Education," *J. Eng. Educ.*, vol. 103, no. 1, pp. 45–76, Jan. 2014, doi: 10.1002/jee.20038.

[10] J. Wigdahl, G. L. Heileman, A. Slim, and C. T. Abdallah, "Curricular Efficiency: What Role Does It Play in Student Success?," presented at the 2014 ASEE Annual Conference & Exposition, Jun. 2014, p. 24.344.1-24.344.12. Accessed: Jan. 20, 2023. [Online]. Available: https://peer.asee.org/curricular-efficiency-what-role-does-it-play-in-student-success

[11] D. Torres, J. Crichigno, and C. Sanchez, “Assessing curriculum efficiency through Monte Carlo simulation,” *J. Coll. Stud. Retent. Res. Theory Pract.*, vol. 22, no. 4, pp. 597–610, 2021.
[12] D. Waller, “Organizational factors and engineering student persistence,” Dissertation, Purdue University, 2022. [Online]. Available: https://doi.org/10.25394/PGS.21606342.v1
[13] R. Layton, R. Long, M. Ohland, M. Orr, and S. Lord, “midfieldr: Tools and methods for working with MIDFIELD data in ‘R.’” 2022. [Online]. Available: https://midfieldr.github.io/midfieldr/
[14] G. Csárdi, “Package‘igraph.’” 2021.
[15] G. Heileman, M. Hickman, A. Slim, and C. Abdallah, “Characterizing the Complexity of Curricular Patterns in Engineering Programs,” in *2017 ASEE Annual Conference & Exposition Proceedings*, Columbus, Ohio: ASEE Conferences, Jun. 2017, p. 28029. doi: 10.18260/1-2--28029.
[16] OpenAI *et al.*, “GPT-4 Technical Report.” arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2303.08774.
[17] 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.
[18] H. Vasquez, A. A. Fuentes, and J. A. Kypuros, “Enriched student guidance and engagement in lower level engineering gatekeeper courses,” in *2016 IEEE Frontiers in Education Conference (FIE)*, Oct. 2016, pp. 1–8. doi: 10.1109/FIE.2016.7757663.