2018 ASEE Annual Conference & Exposition
June 24–27, 2018
Salt Palace Convention Center
Salt Lake City, UT

Paper ID #22429

Work in Progress: An Analysis of Correlations in Student Performance in Core Technical Courses at a Large Public Research Institution's Electrical and Computer Engineering Department

Mr. Christopher Robbiano, Colorado State University

Chris Robbiano is currently a PhD student in the Electrical and Computer Engineering department at Colorado State University. He received a BS degree in electrical engineering and a BS degree in physics in 2011, as well as an MS in electrical engineering in 2017 from Colorado State University. His current areas of interest are statistical signal processing and engineering education.

Dr. Anthony A. Maciejewski, Colorado State University

Anthony A. Maciejewski received the BS, MS, and PhD degrees in electrical engineering from Ohio State University, Columbus in 1982, 1984, and 1987, respectively. From 1988 to 2001, he was a professor of electrical and computer engineering at Purdue University, West Lafayette. He is currently a professor and head of the Department of Electrical and Computer Engineering at Colorado State University. He is a fellow of IEEE. A complete vita is available at: http://www.engr.colostate.edu/ ˜aam.

Prof. Edwin K. P. Chong Ph.D., Colorado State University

See edwinchong.us

Work in Progress: An Analysis of Correlations in Student Performance in Core Technical Courses at a Large Public Research Institution's Electrical and Computer Engineering Department

Introduction

The National Science Foundation is supporting our Electrical and Computer Engineering (ECE) department at Colorado State University (CSU) through their "Revolutionizing Engineering and Computer Science Department" (RED) program. As the focus of this project, we propose to remove the artificial barrier that a traditional course-based curriculum creates [1]. To aide in doing so, we seek to understand the relationships of student performance between technical courses within the ECE curriculum. In particular, we begin by studying the performance between the three core junior level topics, i.e., electronics, electromagnetics, and signals and systems, each spanning two semesters.

As part of the introductory phase of the RED project at CSU, the junior year ECE courses have experienced the beginnings of an overhaul to the methods with which material is presented. The data that is analyzed in this paper comes solely from CSU undergraduate students, but is also available as part of the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) dataset. As part of the MIDFIELD data collection, anonymized student records are kept for every semester that students are enrolled, and include information such as individual course grades, cumulative GPA as well as high school academic records. More information about what information is collected for the MIDFIELD dataset can be found at [2]. To align with the current work in our RED project, we chose to examine only the individual course grades and cumulative GPA from the junior year in this paper.

We used two methods for performing quantitative analysis on the relationships between the first and second semester offerings in each topic as well as the relationships between each topic and the cumulative GPA of the students. The first method is a correlation analysis using the Spearman correlation coefficient as a metric of monotonic association between performance in each course while the second method is a Principal Component Analysis (PCA) to identify the components that contribute the largest amount of variance to the overall performance of students.

In this paper we present the analysis of correlations between individual course grades and a PCA on the course grade data with the goal of identifying notable relationships between the grades and performance between prerequisite first-semester, junior-year courses and their second semester requisite counterparts. The remainder of this paper will be presented in the following format. The literature review section will discuss similar research in the area of identifying relationships between prerequisite courses and their requisite counterparts as well as what are good indicators for future performance. The data section will present the dataset that was used in the analysis, including the choices made to select a subset of the entire dataset. The methods section will present an overview of the Spearman correlation coefficient and an example of calculating the ranked version of a variable as well as an overview of the Principal Component Analysis theory. The results and discussion section will present the relationships that were found within the data using the aforementioned analysis methods. Finally, the future work and conclusion sections will present goals for future analysis between similar datasets and recap the analysis and findings presented in this paper.

Literature Review

There have been a number of similar studies that look to identify meaningful relationships between prerequisite courses and their requisite counterparts as well as identifying good indicators of performance in future courses. Of these studies, many of them tend to include the high school performance of students, including their SAT and ACT scores [3], [4]. In particular, Johnson and Kuennen used a linear regression model to show that in an introductory economic and business statistics course the largest contributing factor to success among gender, race, ACT score, GPA, and others is GPA. They conclude that the GPA going into the course attributes 40.5% of the total influence on the course outcome, and vastly outweighs the influence of any other single indicator [5].

Simpson and Fernandez performed research on a dataset that is similar to the one presented in this paper to identify if strong correlations exist between students' grades during their early semesters and their corresponding performance in mid-level engineering courses. They used a linear correlation instead of a monotonic correlation, as presented in this paper, but found the same conclusion of prerequisite course performance having a strong correlation with the performance in the requisite course [6]. Similar conclusions are found by Hwang, Yu, Su and Tseng [7] in their research on undergraduate students who participated in programming courses, through the use of fuzzy logic association rules. Research done by Easter [8] on undergraduate chemistry students also comes to the same conclusion of prerequisite performance being a strong indicator of performance, through linear correlations of indicators.

Research conducted on larger sets of the MIDFIELD dataset include the identification of relationships between gender, race and trajectory paths of engineering students presented in [9] as well as identifying relationships between a student's ability to graduate college and their performance in both high school and college using monotonic associations [10].

Data

Data records from the ECE undergraduate program, collected over the past 26+ years, were used for this analysis. The full dataset consists of over 2,700 individual student records, but only the student records with a complete set of grades in the junior year courses were included. This reduced the size of the dataset that was used in the analysis to 803 students. Additional statistics about the dataset can be seen in Table 1.

Table 1: Dataset Statistics

Total # of Students	2700	Earliest Record	Fall 1990
Avg. # of Semesters for non-transfer Students	9.87	Latest Record	Spring 2016
% of Students that Graduated in ECE	51.22	% of ECE Majors	100
# of Students with Complete Junior Year Records	803	Distribution of GPA	$\mathcal{N}(2.688, 0.828)$

The distribution of the grades per class can be seen in the violin plot of Figure 1. The multimodal characteristics of each distribution can partially be attributed to the discretization of course GPAs due to conversion from +/- letter grades to their numeric counterpart.

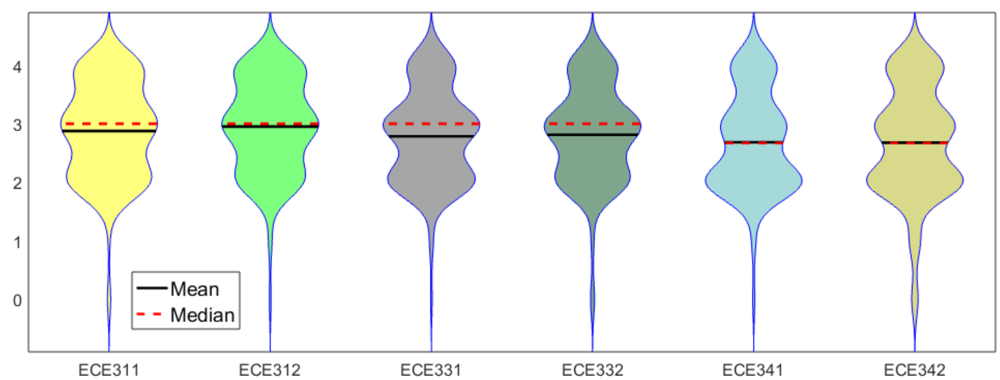

Figure 1: Distribution of Grades per Class

The dataset includes cumulative GPA per semester that is recorded in a 0.00 to 4.00 range, while individual course grades were recorded in a +/- letter grade range from A+ to F. The course grades also include I for incomplete, S for satisfactory and W for withdraw. All +/- letter grades were converted to a range between 0.00 and 4.00 based on Table 2.

In addition to the course grades and cumulative GPA per semester, the dataset contains information such as the location of origin, ethnicity and gender, and previous educational performance if it existed. —- has a "repeat-delete" policy that allows students to retake a course and replace the previous grade with the grade from the latest offering of the course. This means that the most recent grade will be included in the cumulative GPA calculations for the student irrespective of previous performance. The dataset used in this analysis contains the outcomes of *all* course work and thus grades were only used in accordance to the "repeat-delete" policies.

The courses examined are the six major courses presented during the junior year curriculum that cover the three main topics of signals and systems, electronics, and electromagnetics. The course numbers are ECE 311/312, ECE 331/332, ECE 341/342, respectively.

Table 2: Letter Grade to Numerical Grade Conversion

A+/A/S	A-	B+	B	B-	C+	C	C-	D+	D	D-	F/I/W
4.00	3.67	3.33	3.00	2.67	2.33	2.00	1.67	1.33	1.00	0.67	0.00

It is worth noting that in addition to only using complete records, we used Probabilistic PCA to perform infilling of grade data for entries that were missing in student records and the same analysis was performed [11]. The results on the synthetic infilled data was essentially identical to the results of the non-synthetic data and thus we present the analysis on only non-synthetic data.

Methods

We used two different analysis methods in this paper. The first method used is the Spearman correlation coefficient [12], which is able to help identify monotonic relationships between two variables. We chose the Spearman correlation coefficient in favor of the more commonly used Pearson correlation coefficient as we are not necessarily interested in linear relationships between the performance in each course but rather we are interested more generally in whether students will perform better in one course given that they perform better in another course. The second method used is Principal Component Analysis, which in our case is able to help identify the amount of variance in performance that is explained by each individual course. In particular, PCA is able reveal trends in the data that are not easily observable in the raw data [11].

In both methods of analysis, vector $\mathbf{x}_i \in \mathbb{R}^{803\times 1}$ was constructed with data from the ith course where the kth row of $\mathbf{x}_i$ represents the grade that the kth student attained in course i. Each $\mathbf{x}_i$ was normalized by taking its z-score, and a data matrix containing all normalized $\mathbf{x}_i$s was formed such that $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_M\right]$. In the case of the PCA, $M = 6$ as we wish to look at only the variance explained by the course data. In the case of the correlation analysis $M = 7$ with $\mathbf{x}_7$ formed from the GPA data as we wish to examine the associations between courses but also between courses and the overall GPA.

Spearman Correlation Coefficient

Given the general monotonic trends in the data, the Spearman correlation coefficient ρ was used as a measure of association between variables $\mathbf{x}_i$ and $\mathbf{x}_j$ [12]. The Spearman correlation coefficient is a non-parametric measure of monotonic association and operates on the *ranked* version of variables $\mathbf{x}_i$ and $\mathbf{x}_j$. In this context, the process of creating a ranked variable $\mathbf{r}_i$ involves assigning a value of 1 to the smallest element in variable $\mathbf{x}_i$, a value 2 to the second smallest element of $\mathbf{x}_i$, and so on. In the case where N elements share the same value and hence would be assigned the same rank, the rank for each element is calculated by averaging the next available N ranks and assigning each element the averaged rank value. For example, if $\mathbf{x}_{i_4}$, $\mathbf{x}_{i_{13}}$, and $\mathbf{x}_{i_{36}}$ each had the value of 5 and corresponded to a rank of 13, then the rank assigned to $\mathbf{x}_{i_4}$, $\mathbf{x}_{i_{13}}$, and $\mathbf{x}_{i_{36}}$

would be $(13 + 14 + 15)/3 = 14$ and the next available rank would be 16. Each element of $\mathbf{r}_i$ is then assigned the rank of the corresponding element in $\mathbf{x}_i$. A brief example of creating ranked variables is seen in Table 3.

Table 3: Creating Ranked Variables Example

English Exam Score $\mathbf{x}_e$	*Math Exam Score* $\mathbf{x}_m$	*English Exam Rank* $\mathbf{r}_e$	*Math Exam Rank* $\mathbf{r}_m$
100	64	6	1
93	74	5	2.5
68	74	2	2.5
75	88	3.5	5
75	86	3.5	4
60	94	1	6

The Spearman correlation coefficient between variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is described by the following equation, where $\sigma_{\mathbf{r}_i}$ and $\sigma_{\mathbf{r}_j}$ are the standard deviations of ranked variables $\mathbf{r}_i$ and $\mathbf{r}_j$ and $\text{cov}(\mathbf{r}_i, \mathbf{r}_j)$ is the covariance between ranked variables $\mathbf{r}_i$ and $\mathbf{r}_j$ [12]:

$$\rho = \frac{\text{cov}(\mathbf{r}_i, \mathbf{r}_j)}{\sigma_{\mathbf{r}_i}\sigma_{\mathbf{r}_j}} \approx \frac{\frac{1}{n}\sum_{k=1}^{n}\left((\mathbf{r}_{i_k} - \bar{\mathbf{r}}_i)(\mathbf{r}_{j_k} - \bar{\mathbf{r}}_j)\right)}{\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{r}_{i_k} - \bar{\mathbf{r}}_i\right)^2\right)}\sqrt{\left(\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{r}_{j_k} - \bar{\mathbf{r}}_j\right)^2\right)}}.$$

Principal Component Analysis

PCA is a linear orthogonal transformation used for many applications, but primarily dimensionality reduction and explanation of the covariance structure that exists within a dataset [11]. Given the covariance matrix $\mathbf{C}$ of the data matrix $\mathbf{X}$ and its eigendecomposition $\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, PCA transforms data vectors from the original M dimensional space into the principal component space of the same dimension spanned by the columns of $\mathbf{U}$. The ith eigenvector of $\mathbf{C}$ explains $\lambda_i / \sum_{j=1}^{M} \lambda_j\%$ of the total variance of the data in the original space, where λ_i is the ith diagonal element of $\mathbf{\Lambda}$ and the eigenvalue associated with the ith eigenvector $\mathbf{u}_i$. The λ_is are ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$. The set of the first p eigenvectors, $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p\}$, that explain the bulk of the variance in the dataset are referred to as the *principal components*, and in this paper we provide an analysis of the largest principal component. The kth row vector $\mathbf{g}_k$ of the data matrix $\mathbf{X}$ contains the grades for the kth student, and the transformation of each $\mathbf{g}_k$ from the original space to $\tilde{\mathbf{g}}_k$ in the principal component space can be described by

$$\tilde{\mathbf{g}}_k = \mathbf{g}_k\mathbf{U}.$$

Results and Discussion

We present results in this section that help explain relationships between the performance in individual courses as well as relationships between individual courses and the overall GPA. We discuss the correlations between courses and GPA, and we also provide an analysis of the principal component (PC) that explains the largest amount of variance in the dataset.

In the correlation analysis, the highest correlations exist exclusively between the individual courses and the cumulative GPA indicating a strong positive correlation between doing well overall and doing well in each individual course, with each relationship having a $\rho \geq 0.73$. Examining the relationship between individual courses while excluding the GPA, the highest correlations exist exclusively between the sequenced courses that cover the same topic, with each $\rho \geq 0.54$ indicating a moderately strong positive correlation between the performance in the prerequisite and requisite courses. In all cases the results are statistically significant with $p < 0.0001$. The matrix containing all of the Spearman correlation coefficients can be seen in Table 4 and the correlation strength can be seen visually in Figure 2. These correlation results support the intuition that good performance in the prerequisite courses is required for good performance in the requisite courses.

Table 4: Spearman Correlation Coefficient, ρ, between Courses and GPA

	ECE311	ECE312	ECE331	ECE332	ECE341	ECE342	GPA
ECE311	1	0.60	0.44	0.41	0.50	0.52	0.74
ECE312		1	0.47	0.49	0.51	0.53	0.77
ECE331			1	0.54	0.52	0.48	0.73
ECE332				1	0.46	0.48	0.73
ECE341					1	0.56	0.76
ECE342						1	0.78
GPA							1

Note: All entries are statistically significant with $p < 0.0001$

Examining the results of the PCA, the first PC captures 58.18% of the variance described in the data set. The set of PCs and the percentage of total variance explained for each PC can be seen in Table 5. The values in the first PC can be seen to be almost constant across all courses, indicating that there is an almost equal amount of weight provided from each individual course when describing the composition of the first PC.

A *loading plot* was used to help illustrate the connection between the cumulative GPA and the first PC. The loading plot seen in Figure 3 contains the transformed grade vectors $\tilde{\mathbf{g}}_k$s projected onto the subspace defined by the first and second PCs. Each $\tilde{\mathbf{g}}_k$ on the plot has been stylized to reflect the associated cumulative GPA. It can be seen through the distinct banding on the plot that there is an association between the first PC and the cumulative GPA. Investigating this association further, the correlation coefficient was calculated between the first PC and cumulative GPA and was found to be $\rho = 0.9993$ with $p < 0.0001$. This indicates a very strong, almost perfect, positive correlation between the two, and shows that the first PC is effectively a scaled and shifted

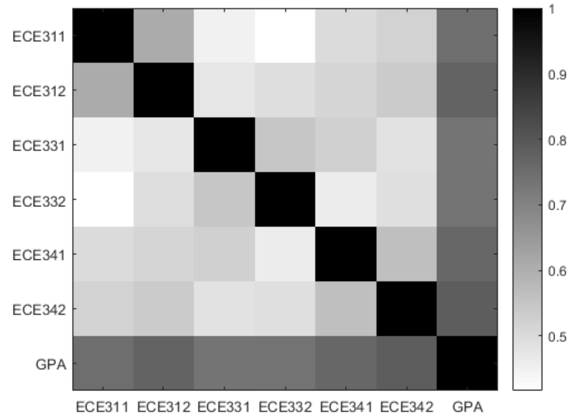

Figure 2: Spearman Correlation Coefficients, ρ, between Courses and GPA

Table 5: PCs and the Percent of Total Variance Explained per PC

	PC1	PC2	PC3	PC4
ECE311	0.4069	-0.5436	0.1150	-0.4333
ECE312	0.4206	-0.3242	-0.3939	-0.2483
ECE331	0.3988	0.4737	0.5312	-0.3404
ECE332	0.3952	0.6008	-0.4948	-0.1040
ECE341	0.4120	-0.0814	0.4948	0.4755
ECE342	0.4150	-0.0849	-0.2441	0.6306
% of Total Variance Explained	58.18	10.38	9.41	8.78

version of the cumulative GPA. Thus, the GPA explains roughly 60% of the total variation in the performance of individual courses.

Future Work

Traditionally, the three topics taught during the junior year are presented without much integration with one another and students commonly have the misconception that each topic is relatively unrelated. This could, in part, be an explanation to the results seen in the examination of the correlations between the individual courses that have shown little inter-topical correlation. The ECE department is currently in the introductory phase of implementation for the Knowledge Integration (KI) portion of the RED project as described in [13]. As part of the KI, seminars are held every four weeks during a semester in which students are exposed to the connections between the topics. In conjunction with each seminar, students complete assignments that mimic real-world engineering problems designed to highlight the intricate dependencies between each topic. As more data is collected in the years going forwards, we will be performing the same analysis that is presented in this paper but on the grade data taken from the time after KIs were

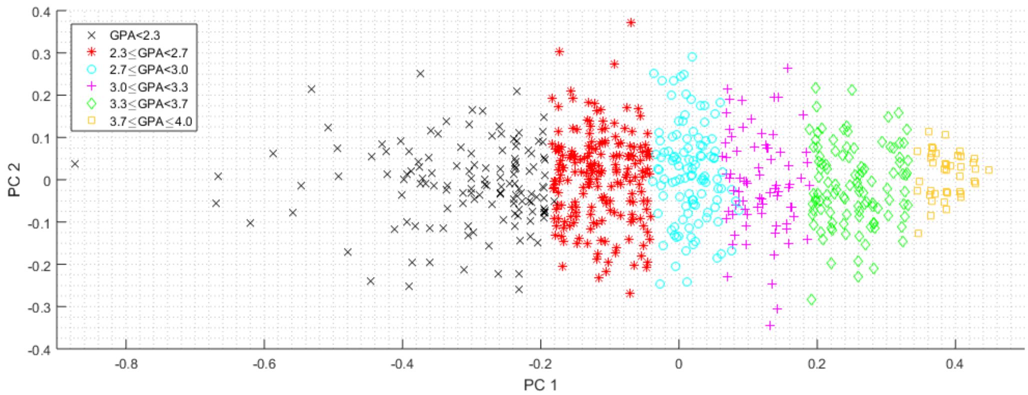

Figure 3: PCA Loading Plot with GPA Annotated

initiated and comparing the results to see if the inter-topical correlations have changed to a more uniform distribution, if at all.

We have performed preliminary work to extend the work presented in this paper to include data from freshman and sophomore level math and engineering courses to help identify further relationships in student performance. We are particularly interested in identifying underlying relationships between topics, such as a student's natural affinity for topics involving multivariate calculus, that may drive students to excel in those particular topics.

Conclusion

In this paper, we presented an analysis of the correlations between two semesters of topical courses as well as an analysis of the dominating principal component generated from the student grade data. The technical results of the correlation analysis and PCA explain two main points. The first is that there is a moderately strong correlation between the performance in prerequisite and requisite courses. The second is that the cumulative GPA explains roughly 60% of the total variance of the performance in individual junior year courses. These two results provide statistical support to the common intuition that working knowledge in the prerequisite courses will provide a notable benefit in the requisite courses, and that students who have performed better in their overall coursework are going to generally do better in their individual courses.

Acknowledgement

This work was supported by the National Science Foundation, IUSE/Professional Formation of Engineers: Revolutionizing Engineering and Computer Science Departments (RED) under Grant EEC-1519438.

References

[1] A. A. Maciejewski, T. W. Chen, Z. S. Byrne, M. A. De Miranda, L. B. S. Mcmeeking, B. M. Notaros, A. Pezeshki, S. Roy, A. M. Leland, M. D. Reese *et al.*, "A holistic approach to transforming undergraduate electrical engineering education," *IEEE Access*, vol. 5, pp. 8148–8161, 2017.

[2] (2018) Multiple-institution database for investigating engineering longitudinal development. [Online]. Available: https://engineering.purdue.edu/MIDFIELD/index.html

[3] S. Huang, "Predictive modeling and analysis of student academic performance in an engineering dynamics course," Ph.D. dissertation, Dept. Engr. and Tech. Ed., Utah State Univ., Logan, UT, 2011, https://digitalcommons.usu.edu/etd/1086.

[4] J. L. Kobrin, B. F. Patterson, E. J. Shaw, K. D. Mattern, and S. M. Barbuti, "Validity of the SAT® for predicting first-year college grade point average. research report no. 2008-5." *College Board*, vol. 5, Jan. 2008, https://files.eric.ed.gov/fulltext/ED563202.pdf.

[5] M. Johnson and E. Kuennen, "Basic math skills and performance in an introductory statistics course," *Journal of Statistics Education*, vol. 14, no. 2, p. null, 2006. [Online]. Available: https://doi.org/10.1080/10691898.2006.11910581

[6] J. Simpson and E. Fernandez, "Student performance in first year, mathematics, and physics courses: Implications for success in the study of electrical and computer engineering," in *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, Oct 2014, pp. 1–4.

[7] C.-S. Hwang, P. Yu, Y.-C. Su, and K.-C. Tseng, "Validation of course prerequisites based on student grade using fuzzy association rules," in *2009 International Conference on Management and Service Science*, Wuhan, China, 2009, pp. 1–3.

[8] D. C. Easter, "Factors influencing student prerequisite preparation for and subsequent performance in college chemistry two: A statistical investigation," *Journal of Chemical Education*, vol. 87, no. 5, pp. 535–540, 2010. [Online]. Available: https://doi.org/10.1021/ed800165t

[9] S. M. Lord, M. W. Ohland, and R. Layton, "Understanding diverse pathways: Disciplinary trajectories of engineering students: Year 3-NSF REE grant 1129383," in *2015 ASEE Annual Conference & Exposition*, no. 10.18260/p.23344. Seattle, Washington: ASEE Conferences, 2015, pp. 26.11.1–26.11.11, https://peer.asee.org/23344.

[10] S. Juan, G. Zhang, Y. Min, M. A Padilla, M. Ohland, and T. J Anderson, "Investigating student outcomes using a longitudinal database and statistical procedures," in *9th International Conference on Engineering Education*, San Juan, PR, Jul. 2006, pp. 20–25, http://www.ineer.org/Events/ICEE2006/papers/3496.pdf.

[11] I. T. Jolliffe, *Principal component analysis*, ser. Springer series in statistics. New York, NY: Springer-Verlag, 1986.

[12] P. Chen and P. Popovich, *Correlation: Parametric and Nonparametric Measures*, ser. Sage University Papers. Sage Publ., 2006.

[13] T. Chen, B. M. Notaros, A. Pezeshki, S. Roy, A. A. Maciejewski, and M. D. Reese, "Work in progress: Knowledge integration to understand why," in *017 ASEE Annual Conference & Exposition*. Columbus, Ohio: ASEE Conferences, June 2017, https://peer.asee.org/29166.