Assessing the Data Analysis Training of Engineering Undergraduates

Mrs. Eunhye Kim, Purdue University, West Lafayette

Eunhye Kim is a Ph.D. student and research assistant in the School of Engineering Education at Purdue University. Her research interests lie in engineering design education, especially for engineering students' entrepreneurial mindsets and multidisciplinary teamwork skills in design and innovation projects. She earned a B.S. in Electronics Engineering and an M.B.A. in South Korea and worked as a hardware development engineer and an IT strategic planner in the industry.

Nathan M. Hicks, Purdue University, West Lafayette

Nathan M. Hicks is a Ph.D. student in Engineering Education at Purdue University. He received his B.S. and M.S. degrees in Materials Science and Engineering at the University of Florida and previously taught high school math, science, and engineering.

Matilde Luz Sanchez-Pena, Purdue University, West Lafayette

Matilde Sanchez-Pena is a Visiting Assistant Professor of Engineering Education at Purdue University. She completed her Ph.D. at the same institution in 2018. Her dissertation explored differences across gender of faculty retention and promotion at research-intensive institutions. Dr. Sanchez-Pena aims to promote a more equitable engineering field, in which students of all backgrounds can acquire the knowledge and skills to achieve their goals. Before engaging in Engineering Education research, she completed graduate degrees in Industrial Engineering and Statistics and contributed to a wide range of research areas including genetic disorders, manufacturing optimization, cancer biomarker detection, and the evaluation of social programs.

WIP: Assessing the data analysis training of engineering undergraduates

Abstract

The need for acquiring data analysis skills has become ubiquitous across many professions. In engineering, this need has been recognized through elements such as the current ABET student outcome 3.b, which states that engineering graduates should have "an ability to design and conduct experiments, as well as to analyze and interpret data." While this outcome is a requirement of engineering programs, the length and depth of the data analysis training of undergraduate students vary significantly across engineering majors, leading some scientists to criticize what they perceive as inadequate quantitative training for engineers. The evolution in the capacity to produce and store data requires an exploration of the current state of data analysis training provided by engineering programs. We begin this exploration through the following research question: What data analysis training has been offered to and acquired by engineering students through undergraduate courses?

In this work, we aim to answer this question through a sequential exploratory mixed methods design. Using the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD), we qualitatively coded records of courses offered to engineering students at one public institution between 1989 and 2011 to generate profiles reflecting different levels of data analysis preparation. These profiles were then quantitatively clustered into six distinct levels. The cluster analysis revealed variable patterns of data analysis preparation across different engineering majors. Results from this study also provide a baseline for employers to evaluate the data analysis training of engineers, especially as it relates to the expanding employment opportunities related to data analysis skills. Further, these results may help to inform potential programmatic evaluations and changes.

Background

During the last three decades, there has been controversy about what data analysis knowledge is required by engineers in order to make sound decisions. An important precedent to the modern ABET criteria asserted that engineers should appreciate five aspects of statistics [1]:

- the omnipresence of variability,
- the use of graphical tools such as histograms, scatterplots and control charts,
- the concepts related to statistical inference,
- the importance and elements of well-planned experimental designs, and
- philosophies of data quality derived from data analysis.

Many have argued that the lack of data analysis skills of engineers has contributed to a loss of innovative competitiveness of the US versus other countries [2]. Some have even blamed significant engineering failures such as Challenger space shuttle explosion on the inability of engineers to perform and interpret basic data analyses [3].

Despite recognition that the needs for data analysis skills have increased due to the expanding capacity to collect and store data, efforts to verify skill development among undergraduate

engineering students have been scarce [4]. This study aims to start uncovering the level of data analysis training provided to undergraduate engineering students by exploring the research question: What data analysis training has been offered to and acquired by engineering students through undergraduate courses?

Methods

Data. For this study, two sources of data were considered. First, the Multiple-Institution Database for Investigating Engineering Longitudinal Development (MIDFIELD) [5] was used to access the information of all courses taken by undergraduate engineering majors at 11 institutions across the United States. Available information included the course codes and the type of degree achieved by all students on record. The version MIDFIELD used for this study included students who graduated from the 11 institutions between 1989 and 2011.

College catalogs served as our second source of data for this study. These catalogues provide academic and programmatic requisites for students entering an institution in a given academic year, administrative requirements, requirements for good-standing and graduation, and other valuable information for current and prospective students. These catalogs also usually include descriptions of the courses offered any given year and recent years' catalogs tend to be published online. The engineering course descriptions provided in these catalogs were, therefore, appropriate to analyze for the inclusion of data analysis instruction. At pilot stage, only one institution was selected for this study based on the greatest availability and interpretability of their college catalog.

Study Design. This pilot study was conducted following a sequential exploratory mixed methods design. The qualitative strand was executed first, through content analysis of all course descriptions in the undergraduate catalogs of the institution under study. This process followed a coding framework based on two elements: a) the different data analysis skills described by ABET's Criterion 3.b, and b) the cognitive levels articulated by each description.

Coding Scheme. In order to limit the space of exploration in the varied engineering curricula, the data analysis skills described by Criterion 3.b were tied to either 1) Laboratory courses or 2) Statistics courses. The first were expected to cover the design and execution of experiments, while the latter were expected to cover skills to analyze and interpret data. While it is acknowledged that these abilities are also acquired and practiced in other contexts, such as senior design or capstone projects, the selected approach of focusing on only laboratory or statistics courses was considered suitable first step for the initial pilot stage.

Catalog descriptions were coded for cognitive level of data analysis content based on Bloom's taxonomy [6], with demonstration of understanding coded as 1, application coded as 2, and analysis coded as 3. Note that no higher levels of Bloom's taxonomy were identified in any course description. Therefore, these three levels were used to assign a quantitative rank to each course. For example, a Statistical Topics in Electrical Engineering course with the description, "This course examines *the use of* probability and statistical concepts in electrical engineering applications…," was coded as a Statistics course (as opposed to a Laboratory course), and was assigned a cognitive level of 2 (application). Figure 1 illustrates possible coding relationships.

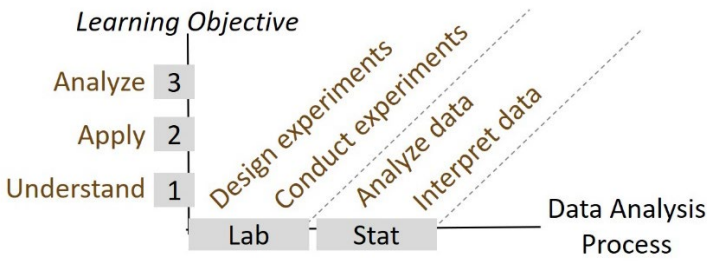

Figure 1. Two-dimensional coding scheme combining data analysis processes and the depth of the learning objectives described by each course.

Once the coding was performed for all Laboratory and Statistics courses using the catalogs, the course and MIDFIELD student data was merged. A weighted average for each type of course was obtained *per student* according to the following calculations:

$$Stats_{score} = \sum_{i=1}^{3} [i \times (\#\ Stats\ courses\ taken\ at\ cognitive\ level\ i)]$$

$$Labs_{Score} = \sum_{i=1}^{3} [i \times (\#\ Lab\ courses\ taken\ at\ cognitive\ level\ i)]$$

Cluster Analysis. Once the two considered scores were obtained, they were used as two-dimensional measures to explore the space where the data analysis training of graduated engineers has been located. Once the two scores were standardized, K-means clustering based on the Euclidean distance on the standardized measures was executed. The total within sum of squares (WSS) per cluster was considered to evaluate the optimal number of clusters to be built. Following the "Elbow method" which graphs WSS for different number of clusters [7], it was identified that 6 clusters was ideal, offering the maximum differentiation between groups and maximum similarity within groups.

Results

The institution used in this study graduated 1,930 engineering students between 1989 and 2011 across six different majors: Chemical, Civil, Computer, Electrical, Industrial, and Mechanical. Demographic characteristics were not considered for this study, but will be explored in future work. K-means clustering was performed using the SPSS statistical software using the previously described two-dimensional representation of statistical training. The cluster distribution is shown in Figure 2.

Observation of Figure 2 shows that the Labs score was more spread than the Stats score. Most groups spanned the majority of the overall Stats score range, suggesting the clustered discriminated primarily by Labs score. When analyzing the composition of these clusters, some were strongly composed of certain majors; for example, cluster 2, which was located in the lowest spectrum of Lab score, comprised mostly industrial engineers, who also had low Stats scores. On the other hand, cluster 4, which was located near the high end of Labs score, was mainly composed of civil engineers. Cluster 6, which was located in the middle and had a

comparatively wide Labs score distribution, was composed primarily of electrical engineers. Cluster 3 was mostly composed of chemical and mechanical engineering students. This group, when compared to cluster 2 had with a similar spread among the Stats score but somewhat higher Lab score. Finally, cluster 1, with the highest Labs scores, was also the second least populated and did not contain a majority of any of the majors, though civil was the most represented. While some of the observed patterns were likely a result of the coding scheme employed, it is clear that differences exist across disciplines.

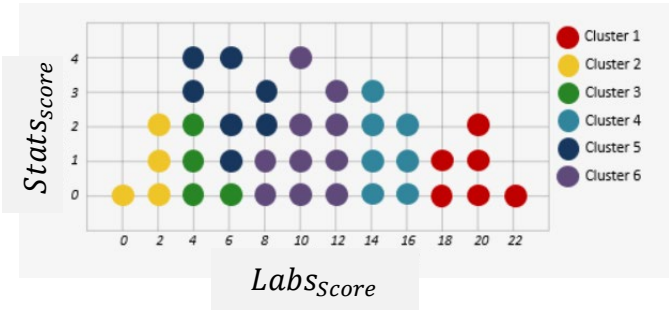

Figure 2. Cluster distribution for Lab and Statistics courses and content at the studied institution.

Table 1. Distribution of clusters by engineering major

	1	2	3	4	5	6	
Chemical	0	4	176	0	0	49	229
Civil	29	3	17	283	0	178	510
Computer	0	1	1	0	0	2	4
Electrical	13	7	22	31	6	480	559
Industrial	0	136	55	1	4	3	199
Mechanical	0	27	373	0	8	21	429
	42	178	644	315	18	733	1930

Discussion and Future Work

The results of this exploratory study provide evidence that, at least at one institution, there have been differences in the degree of data analysis training pursued by students from different engineering majors. However, the differences were more profound for lab-based courses than stats courses. This work's approach can be applied more broadly to evaluate programmatic training of any ABET criteria. In addition, it uses reliable, existing institutional data and allows for additional explorations related to demographic factors such as gender and race.

The next stage of this research includes expanding the number of institutions included. Ideally, all 11 available institutions in the MIDFIELD data will become part of an expanded version of this study. An individual analysis, as well as a collective analysis, would provide more generalizable results about the presented inquiries. Further, while this exercise included only two dimensions, which facilitated interpretation, it is clear that the inclusion of additional dimensions, such as data analyses skills included in other engineering coursework, could provide a more nuanced understanding of curricular elements that promote the development of data analysis skills among engineering graduates. Alternatively, this analysis could incorporate student performance data into the calculation of each score, which would provide a more detailed individual profile by student rather than focusing on the course level only.

References

[1] R. V. Hogg, "Statistical Education for Engineers: An Initial Task Force Report," *Am. Stat.*, vol. 39, no. 3, pp. 168–175, Aug. 1985.

[2] A. Penzias, "Teaching statistics to engineers," *Eur. J. Eng. Educ.*, vol. 15, no. 3, 1990.

[3] F. F. Lighthall, "Launching the Space Shuttle Challenger: disciplinary deficiencies in the analysis of engineering data," *IEEE Trans. Eng. Manag.*, vol. 38, no. 1, pp. 63–74, Feb. 1991.

[4] J. Layton, "Experimental Assessment of Higher-Level Data Analysis Skills," ProQuest Dissertations Publishing, 2013.

[5] M. Ohland, "Engineering persistence studies using the MIDFIELD database," *Esource Coll. Transit. Newsl. Natl. Resour. Cent. First-Year Exp. Stud. Transit.*, vol. 7, no. 4, p. 4, Mar. 2010.

[6] D. R. Krathwohl, "A revision of Bloom's taxonomy: An overview," *Theory Pract.*, vol. 41, no. 4, pp. 212–218, 2002.

[7] B.S. Everitt, S. Landau, M. Leese, D. Stahl. Cluster Analysis, 5th Ed. Wiley & Sons, 2011.