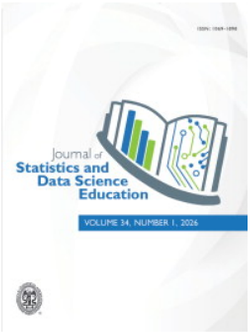


Journal of Statistics and Data Science Education

ISSN: 2693-9169 (Online) Journal homepage: www.tandfonline.com/journals/ujse21

Methodology for Assessing whether a University Provides Students with Opportunities to Learn Data Skills*

Anna Bargagliotti, Robert Rovetti, Suzanne Larson, Thomas Zachariah & Ben Fitzpatrick

To cite this article: Anna Bargagliotti, Robert Rovetti, Suzanne Larson, Thomas Zachariah & Ben Fitzpatrick (07 Jan 2026): Methodology for Assessing whether a University Provides Students with Opportunities to Learn Data Skills*, Journal of Statistics and Data Science Education, DOI: 10.1080/26939169.2025.2610954

To link to this article: https://doi.org/10.1080/26939169.2025.2610954

Methodology for Assessing whether a University Provides Students with Opportunities to Learn Data Skills*

Anna Bargagliotti#, Robert Rovetti, Suzanne Larson, Thomas Zachariah, Ben Fitzpatrick

Department of Mathematics, Statistics and Data Science, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90245

*This work is supported by the National Science Foundation grant no. 1712296

#Anna.Bargagliotti@lmu.edu

Abstract

Due to workforce demands and changing societal needs, ideally *all* university students, regardless of major, should have the opportunity to gain some level of data acumen in their university career. This paper provides thoughtful methods of analysis for measuring how a higher education institution serves its students in providing opportunities to achieve three different levels of data acumen. Using one institution as a case study, the results show that overall students need more exposure to data analysis skills. The methods presented can be generalized and applied to other institutions worldwide thus contributing to the literature on how to assess students' attainment of data acumen in their undergraduate experiences.

Key Words: Statistics Education, Data Science Education, Undergraduate Curriculum

1. Introduction

The demand for people educated in statistics and data science has grown tremendously over the past decade. Jobs related to statistics and data science are expected to grow by about 35% between 2021 and 2031 (BLS, 2023). Moreover, new foundational skills of the digital age for *all* jobs include working with data and making data-driven decisions (BHEF, 2018). Key attributes to be successful in jobs are good computing, analytic and statistical skills, good communication skills, ability to work with real data, ability to tell a story with data both verbally and visually, and the ability to work as a team (Davenport and Patil, 2012, Zorn et al., 2014, Holdren & Lander, 2012). While it is clear that today's workforce demands require increased data acumen (National Academies of Sciences, Engineering, and Medicine, 2018), general societal demands also require more data skills (Gould, 2010,) Data skills are needed to digest and read the news (Engel, 2017), to determine whether information is fake or real (Mihailidis & Viotty, 2017), to make data-driven decisions with health, and to interact with technology online or through devices such as apps, cell phones, and padlets that are collecting data on all of our habits every day (McGrew et al, 2017).

Because of these needs, it is important that students in all disciplines graduate from college with at least some level of data acumen. While there has been clear agreement that students in STEM need to develop data skills, we contend that some level of data skills is needed for *all* undergraduate students, regardless of their major, merely to serve as engaged participants in society today. In this paper, we define three levels of data literacy: Basic, Proficient, and Advanced. This paper then presents a methodology for measuring whether a university provides students with the opportunity to achieve these different levels of data literacy. Our goal is to illustrate a process institutions can follow to measure the opportunities to learn students have at their institutions. As universities aim to better prepare their students for workforce and societal demands, it is not clear whether adequate opportunities exist within university structures for students to deepen their data acumen as they progress through their education. Our methodology provides a systematic and rigorous way to assess the opportunities of a student within a university.

2. Background

Universities in the United States typically have statistics course offerings in several different departments across campus. Because it is very common to have statistics courses housed in different disciplines (e.g., mathematics, computer science, psychology, economics), the American Statistical Association (ASA) and Mathematical Association of America (MAA) have offered guidelines for teaching introductory statistics targeted at non-statistics departments (ASA/MAA Joint Statement, 2014). In addition, several documents published by other disciplinary-specific professional societies describe the statistical and data-related understanding necessary for students within the discipline (e.g., American Sociology Association/ASA, American Economics Association). Often times topics included in courses offered by various departments overlap and yet their prerequisite structures do not allow a student to move from a statistics course offered in one department to a more advanced course offered by another department, thus potentially hindering a student from progressing their data skills. Departments, often rightfully, argue that the type of statistical techniques needed are discipline specific and thus necessitate the offering of a course within a specific discipline. Because data skills are not necessarily gained solely in statistics majors, it is therefore important to understand the full picture of where students have access to data-analytic skills in an undergraduate education. It is particularly salient that the promotion of diversity is kept under close watch as accessing opportunities to gain data-analytic skills in undergraduate education promotes subsequent success in the work-force. Although specific statistical and data analysis techniques do vary from discipline to discipline, certain basic themes of working with data should be present in all courses (Bargagliotti et. al, 2020). Moreover, it is suggested that statistics instruction focus on interpretation and understanding instead of the mere application of formulas (Carver et al., 2016)

3. Learning Outcomes

To begin measuring the opportunities students have at a university to achieve data acumen, an inventory of courses being offered across departments at an institution must be taken. Once such classes are identified, then one must understand the data skills being targeted in each class and how the classes fit together within a university prerequisite structure. Learning outcomes are a popular and useful manner in which universities set goals for student learning within courses. A university may use pre-determined learning outcomes or make learning outcomes of their own to help assess student opportunities. In this paper, we will illustrate the developed methodology using a set of 13 learning outcomes (LOs) for undergraduate statistical literacy (see Table 1) defined in Bargagliotti et. al (2020). The authors developed these LOs using data from a survey and cross-disciplinary working group discussions. They subsequently validated the learning outcomes using a community survey and study.

Table 1. Learning Outcomes for Undergraduate Data Literacy

Learning Outcomes	Description
1	*(Univariate Statistical Process)* Students formulate and/or address questions about univariate data, collect/consider univariate data, analyze univariate data, and interpret results
2	*(Descriptive Measures)* Students can calculate and interpret descriptive measures for quantitative and/or categorical variables to describe characteristics of the data
3	*(Data Visualization)* Students create and interpret basic data visualizations for quantitative and categorical variables
4	*(Basic Inferential)* Students can carry out, and interpret basic inferential statistical procedures for one or two samples
5	*(Bivariate Processes)* Students can carry out, and interpret results from estimating statistical models for bivariate data (e.g., linear regression, interpolation, extrapolation, predictive inference)
6	*(Data-driven Projects)* Students carry out and communicate results from extensive data-driven project(s) related to a real-life problem (extensive is defined as having a single project that takes more than two weeks to complete or a series of projects that take more than two weeks to complete and are worth at least 25% of the final grade)
7	*(Communication)* Students communicate their analyses and the interpretations of their results in a manner that is appropriate to their discipline in the context of the data (e.g., communication could be emphasized with presentations, oral explanations of results, oral/written answers for in-class work, written explanation of results)
8	*(Design and Limitations)* Students can explain the implications of study design, can select appropriate statistical methods for data analysis, and can explain limitations of their analyses and interpretations
9	*(Critical Consumers)* Students become critical consumers of statistically-based results reported in popular media, recognizing whether reported results reasonably follow from the study and analysis conducted
10	*(Multivariate Statistical Process)* Students formulate and/or address questions about multivariate data, collect/consider multivariate data, analyze multivariate data, and interpret results
11	*(Software)* Students use current statistical software or statistical packages that are appropriate to the discipline and context beyond basic Excel or a calculator
12	*(Programming)* Students write a program (using a programming language) to analyze data or extract information from the data
13	*(Advanced Methods)* Students study at least one type of advanced data-analytic methods such as (not limited to): generalized linear models, Bayesian analysis, advanced probability theory and stochastic processes, non-linear models, machine learning, advanced study-design, big data analysis, econometrics, or statistical computing

*Italicized descriptions represent brief labels that will be used to describe the learning outcome throughout the paper. For example, learning outcome 1 will be referred to as LO 1 (Univariate Statistical Process)

Using these 13 LOs, we define three levels of data literacy. Bargagliotti et al. (2020) used the LOs specified in Table 1 to define Advanced data literacy for undergraduate students. Advanced Data Literacy refers to the completion of all of the 13 learning outcomes. An undergraduate completing this level would be deemed as an undergraduate expert with a high level of data acumen. In this paper, we define two additional levels of data literacy: Basic data literacy and Proficient data literacy.

Basic data literacy is defined by learning outcomes 1, 2, 3, 7, 8, and 9 as specified in Table 2. These learning outcomes focus on the statistical process, descriptives, visualizations, communication, design and limitations, and being critical consumers. The Basic level requirements to represent the learning outcomes that hopefully every undergraduate student graduating should achieve. We argue that every student in today's society must have some level of data acumen. The Basic level focuses specifically on skills needed for existing in society today.

We also define an intermediate level, Proficient data literacy, that builds upon the Basic level yet does not aim to meet the Advanced level. Proficient is defined by additional LOs of 4, 10, and 11. These three LOs refer to basic inferential statistics, multivariate questions and problems and introduction of software (see Table 2). The Proficient level is meant to be above a level that all undergraduates should achieve and introduces a degree of specialization. It might be expected that students in STEM disciplines or business might be likely to aim for a completion of the Proficient level.

Table 2. Definitions of Data Literacy Levels

Data Literacy	**Learning Outcomes Met**
Basic	1, 2, 3, 7, 8, 9 LO 1 Univariate Statistical Process, LO2 Descriptive Measures, LO3 Data Visualization, LO 7 Communication, LO 8 Design and Limitations, LO 9 Critical Consumers
Proficient	1, 2, 3, 4, 5, 7, 8, 9, 10, 11 LO 1 Univariate Statistical Process, LO2 Descriptive Measures, LO3 Data Visualization, LO 4 Basic Inferential, LO 5 Bivariate Processes, LO 7 Communication, LO 8 Design and Limitations, LO 9 Critical Consumers, LO 10 Multivariate Statistical Process, LO 11 Software
Advanced	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 (all of the LOs)

The definitions of levels of data acumen may vary as an institution chooses their goals. The levels defined in this paper are merely exemplary so that the proposed methodology can be illustrated.

4. Coursework

To show an example of the proposed methodology at work, we use the courses at a mid-sized private institution in the western part of the United States. At this institution, twenty-nine different undergraduate courses dealing with statistics and data analysis were identified spanning the departments of African American Studies, Biology, Civil Engineering, Economics, Electrical Engineering and Computer Science, Mathematics, Mechanical Engineering, Political Science, Psychology, Sociology, the College of Business Administration, and the School of Education. The courses were identified by interviewing faculty, department chairs, and deans across the university. In addition, course descriptions were combed and course syllabi were consulted. Table 3 shows all the courses and the number of prerequisites each course has.

Table 3. Course Descriptives

Liberal Arts College		
Department	Course Title	Number of Prerequisites
AFAM	African American Studies Research Methods	0
ECON	Introductory Statistics	1
ECON	Accelerated Intro Statistics	0
ECON	Econometrics	3
ECON	Advanced Econometrics	1
POLS	Empirical Approaches	1
POLS	Advanced Empirical Methods	1
POLS	Public Policy Analysis	0
PSYC	Statistical Methods for Psych	1
PSYC	Research Methods	2
PSYC	Advanced Research Methods	4
SOCL	Quantitative Research Methods	0
Department	Course Title	Number of Prerequisites
AIMS	Database Management Systems	2
AIMS	Analytics & Business Intelligence	1
AIMS	*SS:Intro to Big Data & Data Science	0
BADM	Analytical Concepts/Methods for Business	0
Department	Course Title	Number of Prerequisites
EDES	Research Methods/Early Childhood Assessment	0
College of Science and Engineering		
Department	Course Title	Number of Prerequisites
BIOL	Biological Databases	0
BIOL	*SS: Biostatistical Analysis	0
CMSI	Biological Databases	0
CMSI	Artificial Intelligence	2
MATH	Elementary Statistics	0
MATH	Applied Statistics	1

MATH	Intro Probability/Stats	1
MATH	*SS: Biostatistical Analysis	0
MATH	Adv Topics in Probability & Stats	2
MATH	*SS: Stats & Modeling for Teachers	0
MATH	*Probability & Statistics Lab	0
MECH	*SS: Statistical Design of Experiments	0

Source. Data retrieved from LMU Registrar's Office, Upper Division in Dark Grey, Lower Division in light Gray, *Special Topics

AFAM = African American Studies, ECON = Economics, POLS = Political Science, PSYC = Psychology, SOCL = Sociology, AIMS = Applied Information Management Systems, EDES = Education, BIOL = Biology, CSMI = Computer Science, MATH = Mathematics, MECH = Mechanical Engineering

Some courses listed in the table have multiple course numbers due to cross-listings or changing courses numbers during the study period

Using the LOs in Table 1, each course was coded according to whether it covered a LO or not. Therefore, 13 dichotomous variables were created denoting whether or not a course met each LO. For example, introductory statistics (calculus based) offered in the mathematics department was flagged with LO 1, LO 2, LO 3, LO 7, and LO 8 below. The flagging of the courses followed a rigorous process described in Appendix A.

The courses being offered cover a total of 11 different departments and therefore have a wide reach across the university. As shown in the Table 3, The College of Liberal Arts offers 12 courses related to data; the College of Science and Engineering offers 12, the College of Business Administration offers four, and the School of Education offers one course. Of the 29 courses, 9 are lower division (shown in light gray) courses and 20 are upper division courses (shown in dark grey). Of these upper division courses, five were special reading courses offered in small settings. Fifteen courses do not have any prerequisites, eight have one prerequisite, four have two, one course has three, and one course has four prerequisites. The high number of courses with none or one prerequisite suggests high duplication of content at the lower level entry courses and few offerings at the upper division advanced level.

The economics and mathematics departments, not surprisingly, offer the most courses related to data analysis at the institution; economics offers four and mathematics offers seven courses. However, of the seven mathematics courses, two of the courses are special topics courses offered to students wanting to be secondary teachers. The number of courses meeting different learning outcomes is not uniformly distributed across lower division and upper division courses, nor is it uniformly distributed across learning outcomes. Upper and lower division courses can be distinguished by course numbering. A lower division course may be numbered between 100-299 while an upper division course may be numbered 300-599. Table 4 shows the number of courses at each level by learning outcome. We see that LO 1 – LO 5 are primarily satisfied in the lower division courses while LO 12 and 13 are only satisfied in upper division courses. Few courses satisfy LO 9, LO 12, and LO 13; only eight courses meet LO 9, six courses meet LO 12 and seven courses meet LO 13.

Table 4. Number of Courses at each Level Fulfilling each LO

Number of courses at each level fulfilling each LO													
Course Levels	LO1	LO2	LO3	LO4	LO5	LO6	LO7	LO8	LO9	LO10	LO11	LO12	LO13
100-	1	1	1	1	0	1	1	0	1	0	0	0	0
200-	8	8	7	8	8	5	6	7	3	5	6	0	0
300-	4	5	3	4	2	4	4	2	1	4	3	2	1
400-	5	6	5	3	6	6	6	5	2	6	5	2	5
500-	2	2	2	2	2	2	2	2	1	2	2	2	2

Only one course meets the Advanced data literacy criteria, five additional courses meet the Proficient criteria, and only one additional course meets the Basic criteria. Table 5 shows the courses meeting each level. As seen in the table, Advanced Econometrics meets all of the LOs for Advanced data literacy.

Table 5. Courses that Meet All LOs for Basic, Proficient, or Advanced Levels

Basic	Proficient	Advanced
Political Science: Public Policy Analysis	Economics: Econometrics	Economics: Advanced Econometrics
	Political Science: Empirical Approaches	
	Political Science: Advanced Empirical Methods	
	Psychology: Statistical Methods for Psychology	
	Sociology: Quantitative Research Methods	

Interestingly, none of the mathematics courses meet the entirety of even the Basic data literacy criteria. The two introductory statistics courses offered by the mathematics department, *Elementary Statistics* and *Applied Statistics*, each miss the Basic literacy criteria by one LO. *Elementary Statistics*, which is taught as a general education course does not meet LO 8 (Design and Limitations), and *Applied Statistics*, which carries a calculus prerequisite, does not meet LO 9 (Critical Consumers).

5. Methodology for Assessing Students' Opportunities to Learn

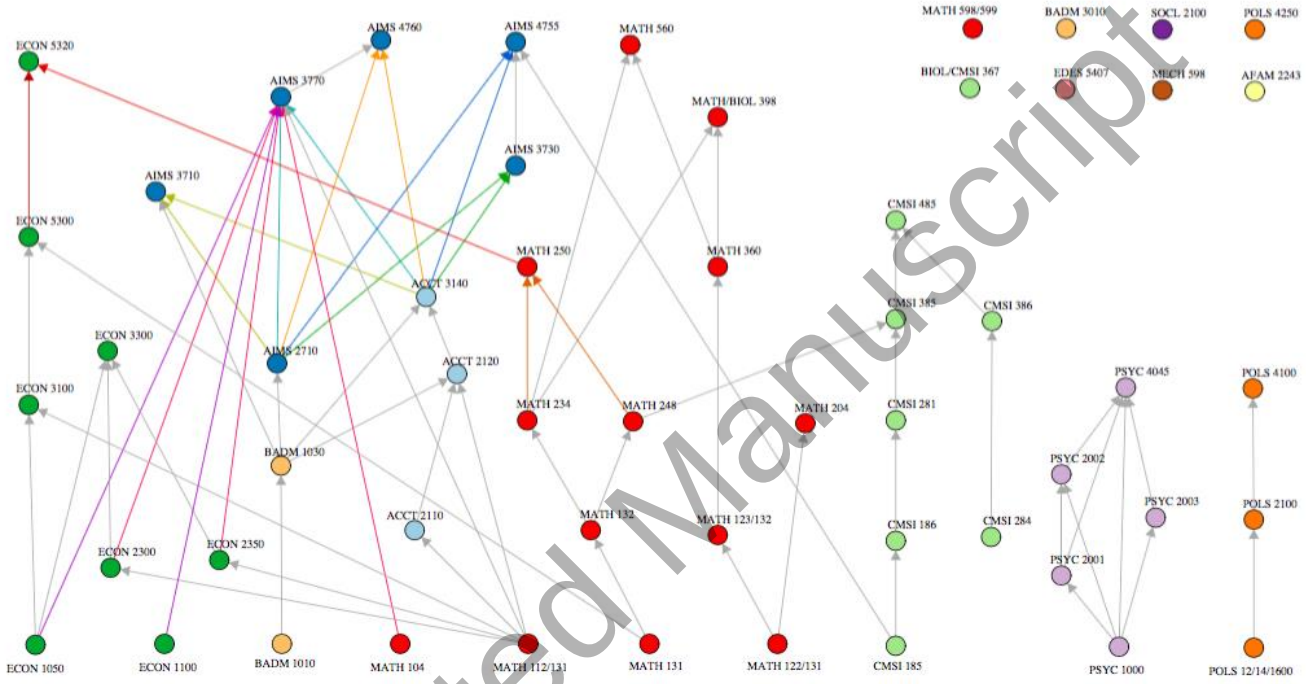
Once courses are identified and learning outcomes are flagged for each course, we then employ an algorithm to determine the possible pathways students can take to achieve the different levels of data literacy and subsequently a cluster analysis to examine similarities and differences of the pathways.

Prerequisite Structure and Pathways

The prerequisite structure of the courses can be studied in order to illustrate which courses and departments are connected to others. Studying the prerequisite structure gives an important snapshot of the possibilities that students have or don't have to access different data-related courses across a university campus. Courses are inherently connected through prerequisite structures in place at the university. For example, several math courses, such as a semester of calculus and a semester of linear algebra, may be required to enroll in an advanced economics course.

Figure 1 illustrates the course dependencies of the 29 courses at the example institution as well as their prerequisites. Each course on the network graph is represented by a colored node on the graph. The color of the node represents the department the course is housed in. Two courses are connected by a directional arrow if a course is a prerequisite for another course. For example, MATH 112 has an outward arrow going to MATH 204 which means that 112 is a prerequisite for 204. While calculus does not contain data analysis, it is represented by a node on the graph because it is a prerequisite to several data courses. If a group of arrows is depicted in color, then an "OR" prerequisite structure is in place, such that any one of those prerequisites is sufficient. For example, AIMS 3770 in blue has a prerequisite of either ECON 3050 or ECON 3100 since those arrows are the same color and point towards the node for AIMS 3770. Otherwise, uncolored arrows indicate an "AND" dependency (all prerequisites to that course must be met).

Figure 1. Network Graph of Prerequisite Structure Among Data-Analysis Courses



Key:

	Course Code	Course Title		Course Code	Course Title
Accounting				ECON 5300	Mathematics for Economics
	ACCT 2110	Financial Accounting		ECON 5320	Advanced Econometrics
	ACCT 2120	Accounting Info for Decision Making	Elementary and Secondary Education		
	ACCT 3140	Accounting Information Systems		EDES 5407	Research Mthds & Early Childhd Assessmt
African American Studies			Mathematics		
	AFAM 2243	African Am Studies Research Methods		MATH /BIOL 398	Biomathematical Modeling
Applied Information Management Systems				MATH 104	Elementary Statistics
	AIMS 2710	Management Information Systems		MATH 112	Calculus for Business
	AIMS 3710	Database Management Systems		MATH 122	Calculus for the Life Sciences I
	AIMS 3730	Programming for Business Applications		MATH 123	Calculus for the Life Sciences II
	AIMS 3770	Production Operations Analysis		MATH 131	Calculus I
	AIMS 4755	Intro to Big Data and Data Science		MATH 132	Calculus II
	AIMS 4760	Analytics and Business Intelligence		MATH 204	Applied Statistics
Business Administration				MATH 234	Calculus III
	BADM 1010	Your Future in Business		MATH 248	Intro to Methods of Proof
	BADM 1030	Bus Perspectives - Info Tech in Organizations		MATH 250	Linear Algebra
	BADM 3010	Analytical Concepts and Mthds for Bus		MATH 360	Intro to Probability and Statistics
Biology/Computer Science				MATH 560	Advanced Topics in Prob/Stats
	BIOL/CMSI 367	Biological Databases		MATH 598/599	Probability & Statistics Lab
Computer Science			Mechanical Engineering		
	CMSI 185	Computer Programming		MECH 598	Statistical Design & Analysis
	CMSI 186	Programming Lab	Political Science		
	CMSI 281	Data Structures		POLS 12/14/1600	US Politics/Comp Politics/Intrntl Rltnshps
	CMSI 284	Computer Systems Organization		POLS 2100	Empirical Approaches
	CMSI 385	Intro to Theory of Computation		POLS 4100	Advanced Empirical Methods
	CMSI 386	Programming Languages		POLS 4250	Public Policy Analysis
	CMSI 485	Artificial Intelligence	Psychology		
Economics				PSYC 1000	General Psychology
	ECON 1050	Introductory Economics		PSYC 2001	Statistical Methods for Psychology
	ECON 1100	Introductory Microeconomics		PSYC 2002	Research Methods
	ECON 2300	Introductory Statistics		PSYC 2003	Brain and Behavior
	ECON 2350	Accelerated Introductory Statistics		PSYC 4045	Advanced Research Methods
	ECON 3100	Intermediate Microeconomics	Sociology		
	ECON 3300	Econometrics		SOCL 2100	Quantitative Research Methods

The network visual immediately reveals that political science (orange), psychology (lavender), African American studies (yellow), education (pink), and sociology (purple) have siloed statistics

offerings. On the other hand, we see that, as expected, mathematics (red nodes) is a key player for prerequisites for economics, computer science, and business. Computer science, economics, mathematics, biology, and business are interconnected; specifically, mathematics is the connector with all of these other departments since mathematics courses are prerequisites for certain courses in these other disciplines (and not the other way around). We also see that there are a total of 8 of the identified data-analysis courses that have no prerequisite attachments (indicated by the singleton nodes on the top right of the graph). This means that these courses can be taken by any student at any time, however, this also means that these courses may not lead a student to more advanced data courses (although they may contribute by satisfying one or more LOs).

Given the prerequisite structures that are in place at the example university, we then find all possible "pathways." See Rovetti & Bargagliotti (2025) for the description and proof of the algorithm used to generate the pathways. We define a *pathway* as a set of courses that meet all of the required LOs per level and that contains all possible prerequisites, even if the prerequisite is not a data analysis course. The pathways must be *maximally efficient*, in that no course can be removed without causing the pathway to no longer meet the requirements of the level LOs. We are interested in finding all possible pathways at the university for reaching all three different levels of data literacy. Once all possible pathways are found, we consider the feasibility of students being able to follow a pathway using a number of metrics and subsequently using clustering methods to identify key critical courses. We begin by defining the total number of individual courses in a given pathway to be the *Pathway Length*.

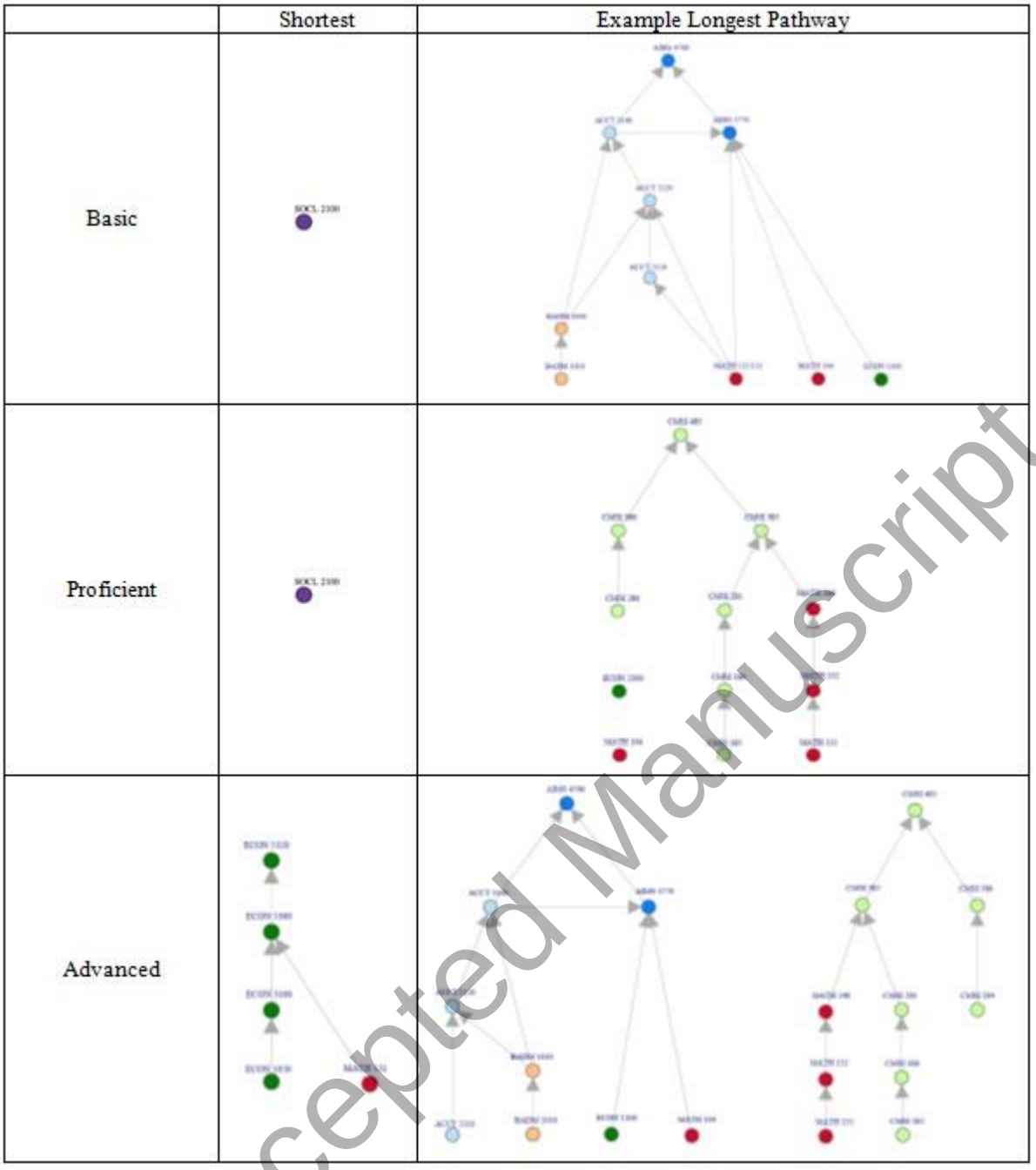
For each of the levels of data literacy, Table 8 shows the number of possible pathways that are available for students to complete as well as the median number of courses within a pathway for each level.

Table 8. Number of Pathways by Level

Data Literacy Level	Total Number of Possible Pathways	Median Number of Courses For Pathways
Basic	22	5
Proficient	40	6
Advanced	104	11

The number of possible pathways is large which is encouraging. Note that the number of possible pathways increases as the levels increase due to the number of ways each Basic pathway can be extended to meet the requirements of the subsequent levels of Proficient and then Advanced. While the number of pathways gives the complete picture of what is possible, many of the pathways are not in fact practical for a student to pursue. We also note that some pathways are minor variations of other pathways, with only the choice of a lower-division, non-LO bearing prerequisite course distinguishing them. Figure 1 illustrates an example of the shortest (least number of courses) and longest (largest number of courses) pathways for each level.

Figure 1. Example Long and Short Pathways for each Level



The shortest pathways for the Basic and Proficient level both consist of only one course. This is because "Sociology 2100", entitled Quantitative Research Methods, meets all of the LOs for both levels. In contrast, the shortest pathway for the Advanced level is given by a five-course sequence consisting of mostly economics classes with one pre-requisite math course. This pathway shows that students in the economics program have a viable pathway towards meeting the Advanced LOs.

One of the longest pathways for Advanced literacy level requires a total of 19 courses in 6 different departments. This is because only a small number of classes exist that meet the higher level LOs and these courses require many prerequisites. Longer pathways may require students to take many extra courses potentially outside of their major. Even at the Basic level, a longest pathway may have 10 courses in it.

Measures of Feasibility. Because pathways may vary widely in length, we further characterize each pathway in terms of the course burden placed upon students enrolled in various academic programs (e.g., economics major, mathematics major, film major). For a given pathway, we then define the *Minimum External Cost* (MinEC) to be the lowest external cost incurred by that pathway over all programs. Similarly, the *Maximum External Cost* (MaxEC) is the highest external cost over all programs. The MinEC and MaxEC represent the best- and worst-case scenarios, respectively, for a given pathway. We also count for each pathway the *Number of Programs with External Cost less than 4* (NPEC4), as a way to measure the general accessibility of that pathway to students across curricular departments. For example, the longest pathway pictured in Figure 2 for basic literacy has a minimum external cost of 2 courses for Accounting majors, and a maximum external cost of 10 courses for all those majors which require none of the 10 courses in that pathway (e.g., History). The longest pathways in the Proficient level in Figure 2, has a minimum external cost of 3 courses for Computer Science majors, and a maximum external cost of 12 courses. The Advanced longest pathway in Figure 2 has a minimum external cost of 10 courses, again for Computer Science majors. These external costs illustrate the difficulty some students may have with access if they are not in certain majors. These analyses provide enlightening insights about how university structures might hinder or help student access to data literacy.

Because students have their own major requirements as well as general education requirements to fill their schedules, it is ambitious to expect a student to take more than 1 to 3 elective courses related to data analysis. This bodes well for achieving the Basic and Proficient levels. Table 10 shows that all majors can achieve Basic and Proficient by taking only one extra course outside of their major (external cost = 1). This is because there is one course, namely Sociology: Quantitative Research Methods, that has no prerequisites and satisfies all Proficient-level LOs, and there are two courses, namely again Sociology: Quantitative Research Methods and Political Science: Public Policy Analysis, that meet all the Basic-level LOs. However, the fact that only one or two specific courses are available to gain these literacy levels at low external cost shows that there is in fact very little flexibility for students (e.g., since the course could not be offered one year, the course could be closed to non-majors). Also, while these courses may meet these particular literacy levels, the subject matter might be far from a students' interest or assume some disciplinary background that is not reflected in the prerequisite. For example, although Political Science: Public Policy Analysis has no prerequisites, it is an upper division course and thus does assume a high level of interest in the specific topic of public policy.

Table 10. Number of Majors per Level for each Amount of External Cost

	External Cost=0	External Cost=1	External Cost=2	External Cost=3
Basic	7	All majors	-	-
Proficient	7	All majors	-	-
Advanced	1	6	11	3

Achieving advanced literacy proves even more difficult. The external costs are high thus showing there are prerequisite university structures that possibly hinder students' ability to become advanced data literate. Only one major can reach Advanced literacy with an external cost of 0 (Economics), six majors can reach Advanced with an external cost of 1, 11 majors with an external cost of 2, and 3 majors with an external cost of 3 extra courses. All 32 other majors at the institution have external cost greater than 3.

Cluster Analysis.

The other component of our proposed methodology relates to identifying the key critical courses in the pathways by applying a standard clustering algorithm to group together pathways based on specific shared course components. Applying clustering methods allows for us to analyze potential

road blocks towards ease of achievement for students towards the three different literacy levels. Clustering approaches group together items (in this case, pathways) based upon some predefined measure of similarity, and both the number of clusters achieved and the notion of similarity is dependent on how fine a distinction between items one wishes to make. Here, the measure of similarity we use is the Jaccard Index (Everitt, Landau & Leese, 2009). Given any two pathways, the Jaccard Index is computed as

$$J = \frac{n_{11}}{n_1}$$

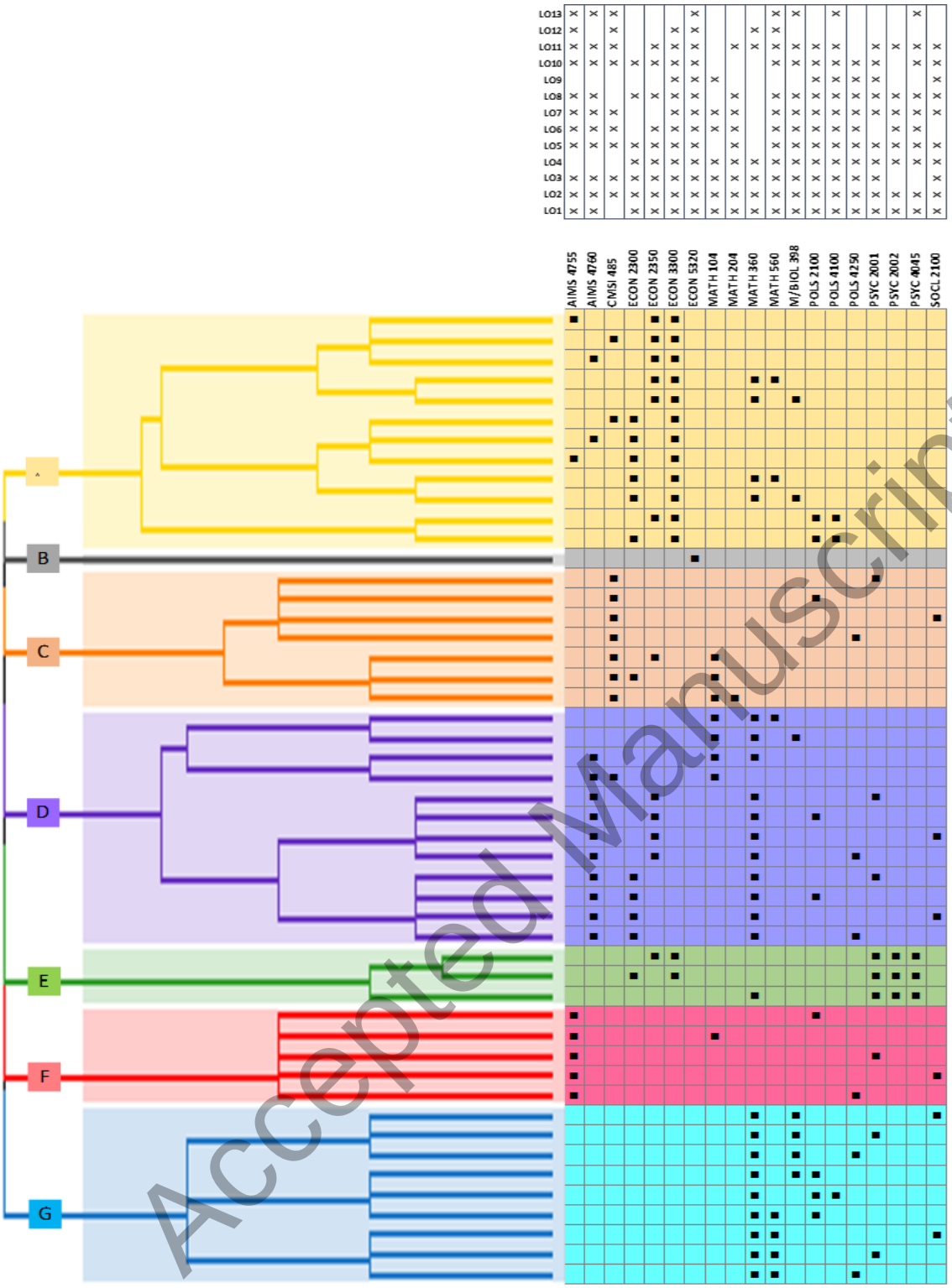
where n_{11} is the number of courses present in both pathways, and n_1 is the number of course present in both pathways combined. Since $n_{11} \leq n_1$ for any two pathways, the Jaccard index ranges from $J = 0$ (the two pathways share no courses in common) and $J = 1$ (the two pathways share all courses in common, and therefore must be the same pathway).

Once these distances are computed between all pairs of pathways, we employ hierarchical clustering to assign each pathway to a distinct group. A number of methods are available to create these groups. We use a bottom-up approach known as "complete linkage". This method operates by iteratively agglomerating two smaller clusters into one larger cluster, if those two clusters are "the most similar" of all existing clusters (see James et al, 2013). The resulting clusters are then examined to determine which critical courses were essential to defining each cluster of pathways.

As an example, we show how this cluster analysis can be used for the Advanced level on all of the possible pathways. It is important to note that an institution utilizing the proposed methodology may choose to limit the pathways considered in the cluster analysis to only those that are feasible or that make sense from a subject and major perspective. Here we use all of the possibly Advanced pathways merely to show the process, however, we recognize that due to the combinatorical nature of counting all of the possibly pathways, there are some pathways that are not feasible for students.

The cluster analysis reveals seven major groupings of pathways each grouping being characterized by the presence of one or two key courses. Each terminal node in Figure 3 depicts one of the 104 Advanced data pathways (pathways that shared the same LO-bearing courses are merged into a single node). Pathways that are located close together by distance within the graph are those that share large number of LO-bearing courses (the measured path lengths along the edges of the graph are directly reflective of the computed Jaccard index values).

Figure 3. Cluster Diagram for Advanced Pathways

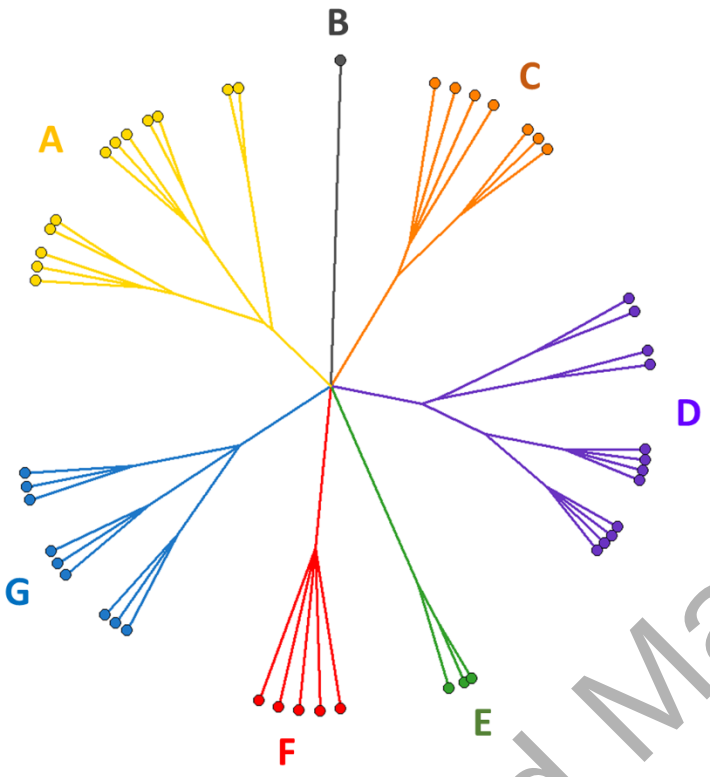


The right side of the Figure 3 contains 49 rows corresponding to the 49 essential pathways, with markers indicating the individual courses contained in each pathway[1]. To the left is a branching structure that encodes the relatively similarities between pathways, as measured by the Jaccard Index. The branching structure can be read much like an evolutionary tree. Pathways which are most

[1] After removing the non-LO bearing courses from the pathways, there remained 49 essential pathways that defined the basis for the clustering. After clustering was completed, each of the original 104 Advanced data pathways were then assigned to their matching cluster.

similar to each other (as measured by the Jaccard Index) are located next to each other, and as the clusters evolve (moving leftwards) we see pathways coalescing into larger and larger groups, until there are seven major clusters remaining. Another way to visually see the clusters is provided in Figure 4. Pathways that share the same color not only are close together and thus have many courses in common but they also are each characterized by specific individual "key" courses that satisfy the Advanced LOs.

Figure 4. Cluster Analysis of Advanced Pathways



Individual courses were identified as "key courses" if they were dominant in a given cluster, and were generally absent in all other clusters. Table 11 gives the characteristics of these key courses color coded by cluster.

Table 11: Characteristics of pathway clusters

Pathway Cluster	Key Upper-division Courses	Average Pathway Length	Average Minimum External Cost (MinEC)	Average Number of Programs with External Cost Less Than 4 (NPECL 4)
A	ECON: Econometrics	9.6	3.2	3.3
B	ECON: Advanced Econometrics	5.0	0.7	9.3
C	CMSI: Artificial Intelligence	11.7	2.7	3.1
D	MATH: Intro to Probability and Statistics AIMS: Business Analytics	12.6	6.0	1.3
E	PSYC: Advanced Research Methods	8.7	4.7	1.0
F	AIMS: Intro to Big Data & Data Science	7.9	2.9	0.8
G	MATH: Intro to Probability and Statistics MATH: Adv Topics in Probability and/or Statistics MATH/BIOL: Biostatistical Analysis	6.4	2.7	2.1

As seen in the table, there are ten courses which serve as "key courses" and define the clusters; these courses derive from five separate programs: Economics, Computer Science, Mathematics, Information Systems, and Psychology. Some of these clusters are more feasible than others in terms of practical completion. For instance, Cluster B (which is defined by the singular presence of Advanced Econometrics) has an average pathway length of 5 courses (due to the four prerequisites needed for Economics 5320). The next-shortest cluster is Cluster G, with an average pathway length of 6.4 courses. Cluster D has the longest average pathway length (12.6) and would not be a practical choice.

The average Minimum External Cost is lowest for Clusters B, C, F, and G, with average values all below 3 courses, indicating that there exists at least one or two majors for whom these are feasible pathways. While the pathway length is high for Cluster C (Computer Science) in particular, its average MinEC is still low, due to the fact that computer science majors would be taking all of those courses as part of their proscribed major.

Because the prerequisites for each key course can generally be assumed to be within the same respective programs, we can draw the general conclusion that students enrolled in Economics, Computer Science, Mathematics, and Advanced Information Management Systems currently have a viable pathway to achieving advanced-level literacy that does not entail taking an undue number of extra courses. We can also note that the 10 key courses are all upper-division courses, most of which are fairly highly specialized and field-centric, which may not appeal to students outside of those programs.

Discussion & Conclusion

This paper demonstrates an in-depth methodology to examine course offerings at a university. The developed algorithms to analyze available pathways and the cluster analysis provide methods for determining key courses, roadblocks, and potential university structures that could hinder or help student access to gaining data acumen.

The results from the case study institution indicate that there are a large number of courses related to data being offered across campus, a total of 29 courses across 11 different departments. While the offerings are encouraging, these courses either miss the mark on some of the important learning objectives necessary in today's statistics education or access to such courses is limited. Access to Advanced data literacy is only really viable to four majors at the university – mathematics, economics, computer science, and business. Because gender imbalances might also be present in these academic programs, from an equity perspective, the access concern is further exacerbated. While achieving Advanced data literacy should be challenging, all students from all major backgrounds should have access to a feasible pathway.

Using the methodologies presented in this study, universities are equipped with a guideline for examining their own curricula. To improve course offerings and student access to data analysis skills, perhaps several courses could also be adjusted to meet specific LOs. For example, at the case study institution, two mathematics courses miss only one LO to classify them as meeting all of the LOs needed for proficiency. Such courses could be purposefully redesigned to ensure that specific LOs are covered that meet the requirements. We found that prerequisite structures hinder student progress towards the different data literacy levels, particularly to the advanced level. Because only a handful of courses across the university meet the difficult LO requirements, there are few pathways that exist that make sense for students. To achieve certain LOs, students incur large external costs. It would not be plausible nor advised for each department to develop their own pathway. This would lead to further siloes and more duplication of efforts across an institution. Instead, a suggestion would be to create an advising group that collaborates on cross-department course offerings and provide each department with a pathway mapping to help advise their students.

It is well documented that data acumen is a worthwhile and needed skills in today's workforce and society. Universities need to address the issue of preparing students to meet these societal needs in a pro-active and effective manner. There is certainly student demand for the acquisition of data skills. For example, recent data shows that from 2016 to 2017, undergraduate degrees earned in statistics grew by 22% to a total of 3,398 (these include both statistics and biostatistics programs) offered across 132 institutions (Pierson, 2018). Our goal as authors is to offer a thought-provoking, systematic, in-depth way to look at undergraduate course offerings. Our hope is that institutions will rigorously evaluate the opportunities and access students have to gain data skills.

Acknowledgments

The authors would like to thank students Nicholas Chew, Amelia Jay, Alison King, and Alexandra MacLean for their work on the tables and figures on this project. In addition, the authors thank the faculty working group consisting of Marie Ford, Natasha Miric, Lance Blaskley, Zaki Eusufzai, Kala Seal, Wendy Binder, and Karen Huchting for thought-provoking conversations regarding the state of data education at the university. Also, we thank Christopher Schmander for his work on organizing the coding sessions and carrying out the reliability analyses. Finally, the authors would like to thank program officer TJ Murphy, ASA's Rebecca Nichols and Donna LaLonde, and the grant advisory board Johanna Hardin and James Albert for their comments on the paper drafts.

Data Availability Statement

The data and code that support the findings of this study are openly available at https://osf.io/fp2xy.

Appendix

In this appendix, we provide details of the methods used to flag the LOs met by each course as well as include a full descriptive table of all the 29 data courses and two large graphics describing the data pathways.

A total of 22 faculty raters participated in the course flagging/coding, with different combinations of raters assigned to each course. Out of the 29 courses, nine were coded by one rater, 16 were coded by two raters, and 4 were coded by three raters. Coding occurred primarily over four Zoom meetings in early April, 2019. The faculty raters who attended each meeting received documents containing the learning outcomes, a set of coding criteria (discussed below), a link to a Box folder containing syllabi for the courses, and a link to a Google spreadsheet where they would enter their codes for their assigned courses. These instructions for coding included the following coding criteria, all of which had to be met for a course to satisfy a given learning outcome:

1. The LO must be covered by course activities and pedagogy, as opposed to simply being a pre-requisite.

2. The LO must be assessed during the course.

 Exams or specific assignments are sufficient for assessing "content" outcomes

 "Process/skills" outcomes must be explicitly emphasized in the class and impact a student's grade, e.g., through a single project, a series of smaller projects, through assignments that require communication, assignments that require the use of software, and so on

 For group work to count as assessment, each person in the group must be required to do work that addresses the outcome.

3. If the course has recently been updated, code it according to the way it is currently being taught.

In addition, faculty were asked to write down their ratings for each course privately before entering them into the Google spreadsheet, to avoid seeing others' ratings before making their own judgments.

After any questions about the coding criteria or process were answered, faculty coded each of their assigned courses according to which learning outcomes the course met, if any, and entered their ratings on the Google spreadsheet. If the faculty raters for a given course were present at the same session, they resolved any disagreements through discussion and entered their final agreed-upon ratings in the spreadsheet, while keeping their original ratings intact for reliability analyses. If the raters for a course attended different coding sessions, they resolved any disagreements later via offline or email discussions. Over the course of the coding session, faculty raters pointed out ambiguities in the coding criteria and learning outcomes. These ambiguities were resolved through discussion before these raters made their independent judgments.

A handful of faculty raters could not make it to any of the coding sessions; these raters received the materials described above via email and coded their assigned courses individually. Once all of the raters had submitted their judgments, reliability analyses were conducted on the ratings they had entered before resolving any disagreements.

The data for the 20 courses that received more than one rating were used to conduct reliability analyses. Notes from the coding sessions and email records post-coding sessions determined that disagreements between raters fell into the following categories:

 Raters interpreted the learning outcomes differently.

 One rater did not fully apply the coding criteria when making his or her independent judgment; upon discussion, the rater in question agreed with the other rater(s) for the course and outcome in question.

 One rater had more specific knowledge about how the course is currently being taught.

 There were differences between raters in how they teach their respective sections of a course.

Table 2 below presents the frequencies of each type of disagreement for each learning outcome.

Table 2. Frequencies of reasons for disagreement between raters for each learning outcome.

Learning Outcome	Reason for disagreement				
	Differences in interpretation of learning outcomes	**Criteria not fully applied**	**More specific knowledge about how course is taught**	**Differences in how sections are taught**	**Not reported**
LO1	1	1	0	1	0
LO2	0	0	0	1	0
LO3	1	1	1	1	0
LO4	0	0	1	0	1
LO5	0	0	1	1	0
LO6	2	2	0	2	0
LO7	3	0	0	0	0
LO8	1	1	0	0	1
LO9	1	0	1	1	0
LO10	3	1	0	0	2
LO11	1	0	0	2	1
LO12	0	1	0	1	1
LO13	0	0	0	0	0

When considering interrater reliability, only the disagreements due to differences in how raters interpreted the learning outcomes are relevant in the context of the study; the other sources of disagreement do not reflect differences in faculty's interpretations of the outcomes per se. Thus, if any disagreements were not due to differences in interpretation, those differences were re-coded to match the final rating agreed upon codes. For example, if two raters initially disagreed about whether a course met LO 1, but then agreed that one rater had more up-to-date knowledge of how the course was being taught, then that rating was re-coded to reflect the up-to-date judgment. If the reason for a disagreement was unknown, the ratings in question were left in disagreement.

Krippendorff's alpha (Krippendorff, 1980) was computed as a measure of interrater reliability for each learning outcome. Compared with other measures of reliability (e.g., Cohen's Kappa), Krippendorff's alpha has the advantage of handling situations with more than two raters, and

situations where raters do not necessarily code each case under consideration – as in the current study, where each course was coded by a maximum of three out of 22 total raters, and most courses were coded by two raters. Krippendorff's alpha penalizes a set of ratings by how often they disagree on specific cases, adjusted for how much disagreement would be expected by chance:

Alpha = 1 – (Disagreement observed/Disagreement expected by chance)

An alpha of 1 represents perfect agreement, with lower values representing lower levels of reliability and 0 representing no reliability, i.e., only a chance level of agreement. While Krippendorff (1980) suggested a value of 0.80 as strong evidence of reliability and 0.67 as a minimally acceptable standard, interpreting alpha ultimately depends on how much uncertainty across raters is acceptable given the goals of a study. Table 3 below presents the alphas for each learning outcome, calculated for the recoded ratings.

Table 3. Krippendorff's alphas for the 20 courses with more than one rater.

Learning Outcome	Alpha
LO1	0.78
LO2	1.00
LO3	0.86
LO4	0.86
LO5	1.00
LO6	0.79
LO7	0.55
LO8*	0.72
LO9	0.79
LO10	0.55
LO11	0.87
LO12	0.71
LO13	1.00

* Only 19 courses received at least two ratings for Learning Outcome 8.

Alpha exceeds the rule of thumb of 0.80 for six of the thirteen outcomes, and falls between 0.67 and 0.80 for another five outcomes. Learning outcomes 7 and 10 each have alphas of 0.55, indicating fairly low reliability. Table 4 below presents the courses where disagreements occurred for each of these two outcomes.

Table 4. Courses for which disagreements occurred on Learning Outcomes 7 and 10.

Course	Disagreement on LO7	Disagreement on LO10
BIOL/MATH 388: Biomathematical Modeling		*
ECON 230/2300: Introductory Statistics	*	*
ECON 235/2350: Accelerated Introductory Statistics	*	*
PSYC 241/243/2001: Statistical Methods for Psychology		*
AIMS 3710: Database Management Systems	*	*

Five out of the 20 courses in the analysis accounted for all of the disagreements that were due to differences in interpretation of outcomes or to unreported reasons.

References

Bargagliotti, A., Binder, W., Blakesley, L., Eusufzai, Z., Fitzpatrick, B., Ford, M., Huchting, K., Larson, S., Miric, N., Rovetti, R., Seal, K., & Zachariah, T. (2020). *Undergraduate learning outcomes for achieving data acumen*. *Journal of Statistics Education, 28*(2), 197–211. https://doi.org/10.1080/10691898.2020.1776653

Business Higher Education Forum (BHEF), (2018). The New Foundational Skills of the Digital Economy. Recovered on September 7, 2019 from: http://www.bhef.com/sites/default/files/BHEF_2018_New_Foundational_Skills.pdf

Bureau of Labor Statistics. (2013). *Employment by Detailed Occupation, 2012 and Projected 2022 Table*, retrieved from https://www.bls.gov/news.release/archives/ecopro_12192013.pdf.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., … & Wood, B. (2016). G*uidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016.* Alexandria, VA: American Statistical Association.

Davenport, T., & Patil, D.J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, 90(10), 70–76.

Engel, J. (2017). *Statistical Literacy for Active Citizenship: A Call for Data Science Education*. Statistics Education Research Journal, 16(1), 44-49. https://doi.org/10.52041/serj.v16i1.213

Everitt, B., Landau, S., Leese, M. (2009). Cluster Analysis 4th Edition. Wiley Publishing.

Farrell, R. & Hussain, M. (2023). Data Occupations with Rapid Employment Growth, Projected 2021–31," *Career Outlook,* U.S. Bureau of Labor Statistics. https://www.bls.gov/careeroutlook/2023/data-on-display/data-occupations.htm

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A pre-K–12 Curriculum Framework*. Alexandria, VA: American Statistical Association.

Gould, R. (2010). *Statistics and the modern student*. *International Statistical Review, 78*(2), 297–315. https://doi.org/10.1111/j.1751-5823.2010.00117.x

Holdren, J. P. and Lander, E. (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

James G., Witten D., Hastie T., Tibshirani R. (2013) Unsupervised Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics, vol 103. Springer, New York, NY.

Mihailidis, P., & Viotty, S. (2017). *Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society.* American Behavioral Scientist, **61**(4), 441–454. https://doi.org/10.1177/0002764217701217

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. The National Academies Press. https://doi.org/10.17226/25104

Pierson, S. (2018). Highlights from 2017 Degree Release: Bachelor's Numbers Close in on Master's. AMSTAT News. August. https://magazine.amstat.org/blog/2018/08/01/2017-degree-report/

Rovetti, R., & Bargagliotti, A., (2025, under review). An Algorithm for Pathways.

Zorn, P., Bailer, J., Braddy, L., Carpenter, J. P., Jaco, W., & Turner, P. (2014). *The INGenIOuS project: Mathematics, statistics, and preparing the 21st century workforce*. Mathematical Association of America.