



Humanizing Data-Driven Methods in Engineering Education Research: A Systematic Literature Review of Four Journals From 2011 to 2021

JIAFU NIU

DAVID REEPING

*Author affiliations can be found in the back matter of this article

LITERATURE REVIEW

VIRGINIA TECH.
PUBLISHING

ABSTRACT

Background: Quantitative methods have been frequently used in engineering education research to investigate generalizable patterns and causal relationships in phenomena of interest to the field. With the proliferation of educational data and growing computational power, researchers are more empowered than ever to wrestle with enduring research questions using machine learning or Big Data methods. However, these techniques can create bias by depressing the voices of individuals belonging to smaller subgroups compared to the majority group in engineering.

Purpose: We examined how engineering education researchers make analytical decisions when they employ data-driven methods. More specifically, we explored how researchers adopted person-centered approaches to offset the inherent issues with data-driven methods.

Scope/Method: We conducted a systematic literature review of the data-driven methods used in quantitative and mixed-method studies published in leading engineering education journals from 2011 to 2021. We used the concepts of person- and variable-centered approaches as a guiding framework to categorize the researchers' analytical decision-making and supplemented the review with a critical perspective using the lens of QuantCrit.

Discussion/Conclusions: Twenty-four articles qualified for the full-text review. Cluster analysis and decision tree models emerged as the sample's two most popular data-driven methods. The findings demonstrated how engineering education researchers adopted person-centered approaches to humanize their chosen data-driven method. This involved finding latent diversity in the sample and understanding the groups formed using the constructs comprising that latent diversity. The coexistence of person- and variable-centeredness in a single study showcased the nuance of leveraging data-driven methods while ensuring the experiences of minoritized populations was not washed out in the noise.

CORRESPONDING AUTHOR:

David Reeping

University of Cincinnati, US

reepindp@ucmail.uc.edu

KEYWORDS:

person-centered approaches;
data-driven methods;
quantitative methods; machine
learning

TO CITE THIS ARTICLE:

Niu, J., & Reeping, D. (2024). Humanizing Data-Driven Methods in Engineering Education Research: A Systematic Literature Review of Four Journals From 2011 to 2021. *Studies in Engineering Education*, 5(2), 150–174. DOI: <https://doi.org/10.21061/see.159>

Engineering education researchers frequently employ quantitative methods to explore causal or correlational relationships. However, according to Godwin et al. (2021), although traditional quantitative approaches can address critical questions in the field, they may also contribute to issues with equity. For instance, when subgroups commonly referred to as “underrepresented” or “minoritized” are combined into a homogeneous category, nuances among them can be buried. Similarly, these methods may exclude these groups as outliers because of their small sample size. Consequently, the findings based on the dominant group might be inaccurately generalized to the entire engineering student population.

Pawley (2017) asserted that research in engineering education often makes broad claims about students, although most of the sample comprises students from the dominant group (i.e., those who are White and male). Pawley emphasized the need for the engineering education research community to acknowledge and address this default whiteness and maleness in their studies. Similarly, Gillborn et al. (2018) argued that statistics have been used to obscure, conceal, and legitimize racism and inequity. They provided an example involving a British government agency and three prominent newspapers, which advanced a narrative portraying White students as victims of racial discrimination. This narrative was supported by statistics showing lower university attendance rates for White students than other ethnic groups aggregated as a single non-White group. However, a more thorough analysis revealed that, individually, White students in Britain constitute the largest proportion of learners entering elite universities and achieving higher grades when compared to each specific minority group. Accordingly, quantitative methods can hold considerable power in educational research depending on how the analyses are framed. This power is amplified with methods that are *data-driven*, both positively and negatively. In this paper, we extend the work of Godwin et al. (2021) by zooming out to a broader scope regarding research design to showcase how person-centered quantitative methods are situated in broader analytical decisions when using data-driven approaches (i.e., those often associated with machine learning).

OPPORTUNITIES AND ISSUES WITH DATA-DRIVEN METHODS

When we discuss data-driven methods, the closest family of techniques that may come to mind is machine learning. Machine learning methods (sometimes cast as educational data mining more broadly) build models by explicitly learning from data that can become adept with prediction and classification tasks. Hence, the data is the mechanism for building the model instead of a researcher testing a presupposed model. Given a large enough dataset, these methods seem to offer a promising means for educational researchers to uncover hidden patterns in service of improving various educational processes (Baker & Yacef 2009).

There are two common types of machine learning approaches—supervised and unsupervised. Supervised machine learning approaches are how people typically perceive such techniques; researchers gather data labeled with some categories (often called classes or class labels) as examples and train the model by showing it these examples. Then, the researchers test the resulting model with data the model has yet to see. Supervised models include decision trees, random forests, Naïve Bayes, neural networks, and other algorithms that try to model the relationship between dependent variables (the classes) and independent variables used during training (Hilbert et al. 2021). The unsupervised machine learning algorithms are exploratory and applied to data where the categories are unknown. The algorithms aim to learn the class labels; typical techniques include cluster analysis, self-organizing maps, and association analysis.

Data-driven approaches, like those offered by machine learning methods, offer a robust framework for analyzing complex, multi-variable, and often non-linear data in ill-defined high-dimensional spaces (Hilbert et al. 2021). In particular, machine learning techniques have become popular in educational research in recent years for addressing a wide variety of purposes, such as developing intelligent tutoring systems, classifying students’ text-based feedback, and predicting students’ academic performance (Korkmaz & Correia 2019). Accordingly, Hilbert et al. (2021) proposed that developing and employing interpretable machine learning techniques (e.g., individual conditional

expectation plots and counterfactual explanations) can help researchers discover the sources of inequalities and make corresponding modifications to reduce the potential unfairness; they also suggest researchers be more cautious on the selected input features which are fed into machine learning algorithms—particularly demographics and other social identities. For example, Kotsiantis (2012) showcased that students' demographics (e.g., gender, marital status, and age) could be used to predict their academic performance. In another study, Kosinski et al. (2013) demonstrated that an individual's sexual orientation, political views, and other personal attributes with some demographics (e.g., ethnic origin) can be easily predicted with great accuracy by just using people's likes on their Facebook account, even though they did not intend on sharing such information. These personal attributes are vulnerable to being taken as sources of discrimination if the models are used for more nefarious purposes (Hilbert et al. 2021). Thus, researchers have been building more sophisticated and fair machine learning algorithms to employ sensitive variables like demographics and other personal attributes in the model and mitigate their discriminatory potential (Kilbertus et al. 2020; Kusner et al. 2017).

Despite the analytical power of employing data-driven methods in educational research to detect and minimize different kinds of inequalities rooted in our educational system, Hilbert et al. (2021) warned that their careless application can reinforce existing biases—leading to algorithmic bias and discrimination. For example, hidden feedback loops, where a system's output is used as an input in future iterations, can potentially perpetuate the extant inequalities when machine learning algorithms are used in contexts that demand more criticality or sensitivity (Kusner & Loftus 2020), especially sub-groups that are vulnerable to the decisions made using the algorithm (Lu et al. 2018), for example, predictive policing. Data-driven approaches adopt a bottom-up framework that prioritizes the relationships embedded within the data itself, devoid of researchers' presumptions—in theory, at least. This contrasts with traditional statistical methods that follow a top-down approach and aim to establish causality between variables and outcomes (Qiu et al. 2018).

The inductive nature of data-driven approaches is often described as “letting the numbers speak for themselves” (Anderson 2008, p. 2). However, this interpretation becomes questionable when we critically reflect on the origins of these numbers. We have control over how we measure, what we measure, and how we analyze the data. Even if the analysis is claimed to be value-free and without a preconceived model, human choices shape the inputs and are guided by normative values. Consequently, Gillborn et al. (2018) caution that the presumed objectivity associated with quantitative research adopting data-driven method(s) carries the risk of reinforcing racist stereotypes and power structures since “numbers are social constructs and likely to embody the dominant (racist) assumptions that shape contemporary society” (p. 173). These issues can manifest in various contexts, as highlighted by Cathy O'Neil (2017) in the book, *Weapons of Math Destruction*, which demonstrates how data-driven methods can be exploited for malicious purposes, both intentionally and unintentionally, within domains such as the justice system, job applications, and online advertising. Echoing Irving and Askeel (2019), ensuring that machine learning models remain firmly aligned with ethical values will present an ongoing and consistent challenge.

In terms of ethical issues, the broader implementation of data-driven methods can be critiqued, including how the data is sourced to build different models. There are open questions about how to make techniques like those used in machine learning more trustworthy (Shokri et al., 2021), and this discussion of the ethics behind data-driven approaches has been even more pronounced with the public release of generative AI systems like ChatGPT (Wach et al. 2023). When the data used to train a model is poorly sourced, whether in terms of quality or stakeholder representation, the results are also impacted. Facial recognition is one of the most common examples of oversight in data collection with data-driven applications, where systems frequently misidentify non-White and non-male individuals because of the large proportion of the training set being White male faces (Waelen 2023). Therefore, researchers must be keenly aware of their dataset's quality and disclose its limitations if issues cannot be addressed.

How researchers source the data for data-driven methods is also a contentious issue, especially regarding how consent was obtained from individuals to use their data for different purposes (Jo & Gebru 2020). One particularly salient example of how researchers sidestepped participant consent is the “Tastes, Ties, and Time” dataset curated from the Facebook profiles of approximately 1700 college students over three timepoints, which included copious amounts of personal information (e.g., home state, sexual interests, and political views) merged with students’ academic records (Zimmer 2010). The students were not given a chance to opt out of data collection or even informed their data had been scraped. Despite attempts to de-identify the data, it was quickly determined that the anonymized college was Harvard University, and the dataset was withdrawn after the researchers faced a wave of criticism for the oversights in human subjects protections. Although the underlying data can be more or less public, like the Facebook profiles from the “Tastes, Ties, and Time” dataset, there is considerable divergence in what individuals perceive as appropriate use of their data by companies and governments (Auxier et al. 2019). For example, in the Pew Research Center survey of Americans in 2019, 41 percent of those surveyed deemed it appropriate for fitness apps to share data with medical researchers. In contrast, only 27 percent agreed with social media companies using post activity and engagement to monitor users for depression. Although data ethics and ensuring the rights of human subjects are upheld are not unique to data-driven approaches, the impact of the models derived from them elevates the concerns—especially considering how the approaches are discussed in the literature.

THE PREMISE OF QUANTCRIT

The aforementioned critiques of data-driven approaches can be associated with the analytical framework of QuantCrit, which draws from Critical Race Theory to examine conventional practices in quantitative research that may have harmful implications. Gillborn et al. (2018) extensively discuss five core principles that encompass QuantCrit: (1) the centrality of racism (recognizing the pervasive nature of racism in society, acknowledging that some scholars argue against its quantifiability); (2) numbers are not neutral (researchers decide what to measure, whom to measure, and how to measure); (3) categories/groups are neither natural nor given (such as race and gender, are socially constructed rather than naturally given); (4) data cannot speak for themselves (data cannot convey meaning on their own and require interpretation); and (5) social justice and orientation (QuantCrit challenges the notion of assumed objectivity and political neutrality in the application of quantitative research). These principles are broad rejections of practices that are more or less accepted in quantitative research (e.g., using race as a predictor in a regression model), and it can be challenging to see practical steps forward for embodying a QuantCrit mindset at times. To make these principles more concrete, the idea of person-centered approaches has been advanced as a path forward for researchers to engage in their work more critically.

PERSON- AND VARIABLE-CENTERED APPROACHES

To explain what a *person-centered approach* constitutes, we draw on work by Woo et al. (2018), who described three distinguishable conceptualizations of the term, *person-centered*, in scientific research. The first use of the term refers to research focused on individual characteristics rather than situational traits. From this point of view, a study qualifies as person-centered when it examines aspects like personality, skills, or people’s abilities. Such a perspective casts the most extensive net on what counts as a person-centered approach to research, perhaps to the extent that entire disciplines in the social sciences could be classified with the person-centered label—for example, the community of psychologists studying personality could be reasonably deemed as entirely person-centered with this first definition. On the other hand, the second use of the term person-centered is to denote research that emphasizes subjective experiences, in contrast to research concentrating on “objective” aspects of individuals (Weiss & Rupp 2011). Here, person-centered approaches seem most aligned with using qualitative approaches, which can capture more of the diversity inherent to peoples’ experiences.

According to Howard and Hoffman (2018), the third use of the term person-centered refers to a set of analytical methods that aim to identify sub-groupings in a sample based on their similarities in the selected variables. Unlike the previous two conceptualizations of person-centeredness, which focus on the constructs under study, this last perspective shifts the attention to the analytical methods in the study. For this research, we adopt the third conceptualization to discuss person-centered approaches because of our interest in exploring how engineering education researchers offset the potential discriminatory power of data-driven methods through their research design.

Because of their ability to provide detailed insights into the study's sample, person-centered approaches have gradually gained popularity over the years (Lanza et al. 2013; McCutcheon 1987; Vermunt & Magidson 2004). Engineering education was first introduced to the concept most directly by Godwin et al. (2021), who offered the community this way of thinking to differentiate between the underlying mechanisms of various quantitative approaches. A person-centered approach aims to acknowledge and understand the heterogeneity among individuals in a sample by identifying latent groupings based on shared measurements that reflect individual characteristics and their environment. Thus, person-centered approaches are especially useful when the sample contains diversity that is not readily identifiable using pre-existing categories; moreover, such approaches can also uncover different developmental trajectories in longitudinal settings (Muthén & Muthén 2000). Authors writing about person-centered approaches tend to identify a handful of specific analytical techniques that accomplish the goal of uncovering latent diversity in a sample, such as latent profile analysis, latent transition analysis, growth mixture models, and cluster analysis (Godwin et al. 2021; Hofmans et al. 2020; Howard & Hoffman 2018; Muthén & Muthén 2000).

Among the conventional statistical methods in authors' collections of person-centered approaches, data-driven techniques like cluster analysis can be seen. Godwin et al. (2021) argued that although data-driven methods may be used to conduct a person-centered analysis, they require great caution and critical thought during implementation. For instance, Principal Component Analysis (PCA) is a data-driven method commonly employed to reduce the dimensionality of a dataset. However, it is not inherently performed in a person-centered way because it consolidates variables into composite quantities (i.e., principal components), potentially obscuring relationships among the original variables by creating new variables that may not be theoretically meaningful or generalizable to other settings, despite its aim to minimize information loss during the transformation.

Researchers are likely more familiar with the parallel concept of person-centered approaches, analyses that are *variable-centered* (Howard & Hoffman 2018). Unlike person-centered approaches, variable-centered approaches presume the homogeneity of individuals in a sample, and every individual in the sample can be represented by the same set of estimated parameters (Morin et al. 2016). These analyses prioritize prediction and relationships among variables (Laursen & Hoff 2006); accordingly, the commonly used analytical methods for variable-centered analysis include standard statistical techniques, such as ANOVA, regression, and factor analysis (Howard & Hoffman 2018).

When conducting empirical research, it is critical to choose a suitable approach based on the research questions (Howard & Hoffman 2018). Thus, we cannot explore variable-centered and person-centered approaches in a vacuum. Person-centered and variable-centered analyses have traditionally been regarded as distinct tasks, each employing unique models and software (Muthén & Muthén 2000). Although person-centered approaches can provide more comprehensive and detailed insights into groupings within a dataset, it does not imply that they should entirely replace variable-centered methods (Godwin et al. 2021). Neither approach is always a better choice than the other (Howard & Hoffman 2018). Instead of viewing variable and person-centered approaches as a binary choice between right and wrong, researchers are encouraged to select quantitative methods based on their research questions and the strengths of different approaches (Laursen & Hoff 2006).

With this new framework for understanding quantitative methods in engineering education, there is an opportunity to reconsider how data-driven approaches are used within the community. Furthermore, given the fresh perspective of person-centered approaches, it remains unclear how extensively engineering education researchers currently employ these techniques and whether they are applied intentionally with a person-centered mindset. By examining the existing literature for examples of thoughtful applications of data-driven methods using person-centered approaches, these studies can serve as models for future endeavors across various research domains within engineering education.

RESEARCH AIMS

This study explores how engineering education researchers use data-driven methods and examines how they incorporate person-centered approaches into their analyses. The primary research question guiding this investigation was: “How do engineering education researchers employ data-driven quantitative methods and humanize them through the use of person-centered approaches?” Specifically, we aimed to gain insights into the person and variable-centeredness of how different data-driven methods are applied in practice. Godwin et al. (2021) suggested an inherent person-centeredness or variable-centeredness for various common methods, like cluster analysis and PCA. However, zooming out to the broader scope of the design can showcase how these methods are situated in the broader analytical decisions made by the researchers—unveiling nuanced procedures in preprocessing, post-processing, and their general application.

METHODS

To delve deeper into the use of data-driven methods and person-centered approaches in engineering education research, we undertook a Systematic Literature Review (SLR). SLRs have been instrumental in establishing an evidence base for informing research, practice, and policy by systematically synthesizing, evaluating, critiquing, and summarizing existing literature on a particular topic (Borrego et al. 2014). Following the established process for conducting an SLR, specifically adhering to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Borrego et al. 2014), we analyzed publications published within the decade of 2011–2021 from four popular engineering education journals. Within each article, we extracted the design procedures—including data collection, analysis, results, and conclusions—to gain insights into how person-centered analyses were integrated within data-driven approaches.

POSITIONALITY OF THE RESEARCHERS

As a Master’s student of engineering education who holds a bachelor’s degree in Applied Statistics, Jiafu, always seeks opportunities to apply his statistical knowledge in engineering education research. More specifically, he has been interested in how those powerful data-driven methods are applied in quantitative educational research and how they handle the associated equity issues brought by using these techniques (e.g., algorithmic discrimination).

As a student from China, oftentimes he doubts the appropriateness of aggregating students from other countries into a single category, “international students,” and generalizing the findings to the entire group, which actually contains many subgroups. For example, language is one of the most significant obstacles faced by many students from other countries when they are studying in the US. However, when international students are treated as a homogeneous group in such research, the inappropriateness is evident because some international students speak English as their native language, but some do not. A project funded by Division D (Measurement and Research Methodologies) of the American Educational Research Association that was funded during his studies, which investigated how engineering education researchers apply data-driven methods and how they adopt person-centered approaches to humanize these methods, perfectly fit his research interests and allowed him to unpack some of the uneasiness he has had about certain practices in quantitative methods as an insider of an underrepresented group.

David supervised Jiafu in this research. David's positionality has been documented extensively in previous publications (Hampton et al. 2021; Hampton & Reeping 2019). Relevant to this effort, David has been reforming his perceptions of quantitative methods since his first introduction to critical quantitative methods, and has been voicing his concerns about quantitative research practices in engineering education. In light of frustration with more nebulous recommendations from the literature for how to engage in critical quantitative methods, his aims to provide more tangible steps for researchers looking to engage in QuantCrit practices and—in a many cases—uncover what practices authors are already doing that align with the principles of critical quantitative methods. Jiafu and David met continuously throughout the project to discuss one another's perceptions of what it means to engage in QuantCrit and person-centered analyses, which involved revisiting the data, rethinking categorizations of approaches, and delving into different components of the reviewed papers. Thus, this collaboration reflects a reflective construction of themes that support person-centered analyses and highlight what practices already exist in publications using data-driven methods.

DATA COLLECTION AND PREPARATION

The initial stage of data collection in an SLR encompasses defining the inclusion criteria, constructing suitable search terms, and selecting relevant databases, which included each journal's website and Education Research Complete database. Given the exploratory nature of our SLR, we did not impose stringent inclusion criteria. The following list outlines our inclusion criteria:

- Must use a quantitative or mixed methods research design
- Must use one or more data-driven methods in their analytical approach
- Must be an empirical study
- Must involve human subjects in the data analysis, meaning the unit of analysis is the individual

We were specifically interested in articles that applied a data-driven technique with human subjects to learn more about the individuals in the study; this means that studies conducted to create an instrument were not considered within the scope of this SLR. For example, the data-driven technique, Exploratory Factor Analysis, is frequently used to find correlated groups of items in an instrument. This kind of study would be excluded because the unit of analysis is not the individual—instead, it is the set of items in the instrument. More generally, a data-driven method is an inductive research approach that does not depend on an a priori theory or model.

To create our search string, we combined terms for common data-driven methods, such as “decision tree” and “principal component analysis,” with “OR” as the conjunction. More generic terms, such as “machine learning,” were also added. The specific methods were appended based on techniques discussed in Tan et al.'s (2005) *Introduction to Data Mining* to include articles that did not make general comments about their approach being data-driven or machine learning. The full search string is provided here:

“cluster analysis” OR “association analysis” OR “item-set” OR “item set” OR “rule based”
 OR “rule-based” OR “classification” OR “random forest” OR “machine learning” OR “data
 mining” OR “principal component” OR “decision tree” OR “KNN” OR “nearest-neighbors”
 OR “nearest neighbors” OR “support vector “ OR “exploratory factor analysis” OR “EFA”
 OR “logistic regression” OR “Bayes”

The articles were collected from four leading engineering education journals: the *Journal of Engineering Education* (JEE), the *European Journal of Engineering Education* (EJEE), the *International Journal of Engineering Education* (IJEE), and *IEEE Transactions on Education* (TOE). We chose the publication period of 2011–2021 to cover the period where data-driven methods were being adopted into more mainstream engineering education research. We retrieved the articles from Education Research Complete and Wiley's search engine for the *Journal of Engineering Education*. The PRISMA flowchart for this study is shown in Figure 1.

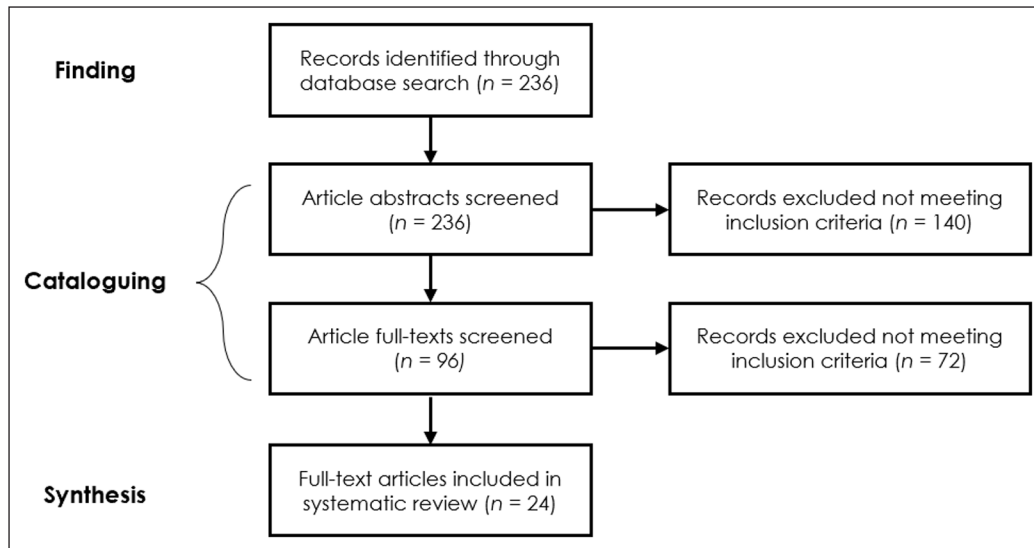


Figure 1 PRISMA diagram of SLR procedures.

A total of 236 articles were retrieved based on the established search string. Following the search phase, we manually screened the abstracts and excluded 140 articles based on the inclusion criteria. Three main reasons led to the removal of articles included (1) the failure to employ a data-driven method (e.g., a data-driven method was mentioned as part of the literature review or future work but not used in the study), (2) the utilization of a data-driven method without involving human subjects, and (3) the absence of an empirical study design. Consequently, 96 articles remained and were categorized based on their primary quantitative methods, including Exploratory Factor Analysis (EFA), Logistic Regression, Cluster Analysis, Principal Component Analysis (PCA), Decision Tree, Random Forest, and Others. A temporary “Others” category was created to accommodate articles where the data-driven methods described did not fit into the aforementioned six categories (e.g., simultaneous use of multiple approaches) or where generic terms like “data mining” or “machine learning” were used to describe the quantitative methods. These articles required further investigation to identify the specific data-driven methods employed by the authors.

Subsequently, we proceeded with a thorough full-text review of the remaining articles to ensure compliance with the inclusion criteria. One significant decision we made was to tentatively exclude the logistic regression group. We observed that articles within the logistic regression category predominantly employed traditional statistical modeling approaches rather than adopting an inductive data-driven framework. We eliminated one article from the PCA group, namely Chan and Fong (2018), because their study discussed the results of PCA from other researchers’ studies rather than directly applying PCA themselves. Furthermore, upon further inspection, Pizard and Vallespir (2017) was also excluded from the Cluster Analysis group since it was later found to not involve human subjects after a closer read. As a result, a total of twenty-four articles remained and advanced to the synthesis stage. Note that we were not specifically interested in making judgments about the quality of the papers we reviewed. Thus, no papers were excluded for issues in quality.

ANALYSIS OF THE SELECTED ARTICLES

To gain a comprehensive understanding of how engineering education researchers incorporate person-centered approaches to humanize data-driven analyses, we employed a three-round coding process. In the initial round, we used descriptive and In Vivo coding (Saldaña 2013) to categorize each article’s methods and results/discussion sections. The initial In Vivo coding allowed us to survey the breadth of how the data-driven approaches were used and establish preliminary patterns. In the second round, the papers were revisited to extract additional context related to the researchers’ purpose for using data-driven method and any additional procedures to delve deeper into their findings. Subsequently, in the third round, we aggregated the findings from the previous coding cycles using descriptive coding to abstract the In Vivo codes to highlight

the person- and variable-centeredness manifested in the application of each technique in different projects. The final round of coding identified tangible strategies for how researchers can balance person- and variable-centered approaches in their designs. To situate the discussion in the broader conversation about the critical application of quantitative methods, we aligned the findings with the QuantCrit framework. This builds off prior work to provide engineering education researchers with practical methods for engaging in a QuantCrit way of conducting quantitative research (Reeping et al. 2023).

RESULTS

To organize our results, we will focus on the person-centeredness observed across the articles and weave in the variable-centered ideas as a comparison. Moreover, we will backdrop the categorization with the corresponding QuantCrit principles manifested from researchers' analytical decisions where applicable. Table 1 shows the distribution of the data-driven methods across the final set of 24 articles as their primary analytical approach. Table 2 summarizes the different elements of the reviewed studies' person- and variable-centered aspects.

PRIMARY DATA-DRIVEN METHOD	NUMBER OF ARTICLES	% SAMPLE
Cluster Analysis	13	54.2%
Decision Tree Methods	4	16.6%
PCA	2	8.3%
Hidden-Markov Chain	2	8.3%
Naïve Bayes	1	4.2%
Topic Modeling	1	4.2%
Model Comparison	1	4.2%
<i>Total</i>	24	

Table 1 Summary of the composition of the final sample of selected articles.

METHOD	PERSON-CENTEREDNESS	VARIABLE-CENTEREDNESS
Cluster Analysis	• Embedding a data-driven technique within a mixed methods study • Embracing and exploring outliers • Allowing fluid membership among groups • Tracking movement between clusters longitudinally • Exploring social dimensions of race as variables	• Cluster analysis as a means of validation • Preprocessing variables using principal component analysis
Decision Tree(s)	• Exploring the branches of a decision tree for latent groups	• Using demographic variables as predictors
Principal Component Analysis		• Preprocessing to form data-specific composite variables to compare across groups • Using principal components for validation
Bayesian Techniques	• Embedding a data-driven technique within a mixed methods study	• Predicting membership to potentially racialized categories • Model comparison focused on accuracy
Topic Modeling	• Embedding a data-driven technique within a mixed methods study	

Table 2 Summary of person and variable-centeredness found in the sample.

One way a researcher could use a data-driven technique as part of a person-centered approach is by embedding it in a mixed methods design to expand upon relationships found in the sample. For example, consider the method *topic modeling*. Topic modeling refers to natural language processing algorithms designed to extract topics from text (Cutumisu & Guo 2019). One of these topic modeling techniques is Latent Dirichlet Allocation (LDA), which is an unsupervised machine learning algorithm assuming a random mixture of hidden topics in different documents in a corpus (Johri et al. 2011). Cunningham-Nelson et al. (2019) leveraged LDA to summarize and visualize student satisfaction from their open-response comments in the end-of-semester course evaluations. Students were asked to rate close-ended questions on a five-point Likert scale, followed by optional free-text responses. Cunningham-Nelson et al. (2019) pointed out that sometimes, the Likert score, as a quantitative value, cannot accurately reflect students' complete thoughts on a given closed-ended item. For example, a student gave a score of 3 to one of the close-ended questions in the survey, which corresponds to a neutral attitude. However, the comments written by this student conveyed a relatively strong negative satisfaction with the content taught in that course. Put another way, the comments written by students were able to contextualize students' satisfactions and struggles embedded in their numeric rating, which provides educators with more informed ideas to improve instruction or curricular design. The scores and open-ended comments occasionally diverged, reflecting the QuantCrit principle, "data cannot speak for themselves." The numbers themselves can be meaningless or misleading depending on how survey items are designed and presented—and to whom they are presented. This application of topic modeling demonstrates how data-driven methods can help uncover a broad range of student experiences or perceptions. This idea also aligns with the premise of *initiation* in mixed methods research (Greene et al. 1989), which involves seeking divergence and reconciling those findings.

Another illustration of how data-driven methods can be embedded in a mixed methods study was Faber and Benson (2017), who employed cluster analysis to identify latent groups using individual students' noncognitive characteristics. In this case, the authors aimed to group students based on student's diverse epistemic motivations and beliefs without examining demographics as a factor in the groupings. After forming the clusters, they sampled students from each group to participate in follow-up interviews. This implementation of cluster analysis aligns with the ideals of person-centered approaches by uncovering latent diversity to describe the sample, and with the QuantCrit principle—categories/groups are neither natural nor given. Faber and Benson's (2017) approach encapsulates the idea of using a clustering algorithm to go beyond immutable characteristics and use qualitative methods to contextualize the sample's latent diversity.

Incorporating qualitative context enhanced their exploration of nuanced differences within subgroups, leading to a deeper understanding of individual responses. This is often associated with the *complementarity* rationale for using a mixed methods design—gaining a more complete understanding of a phenomenon (Greene et al. 1989)—but it can also lead to divergence (Pluye et al. 2009). For example, the study by Ko and Leu (2021) involved binning students into predefined categories (pass and fail) and pitting different machine learning techniques against one another to determine which one had the best predictive power; all seven machine learning algorithms were implemented in variable-centric ways. After testing which algorithm has the most superior predictive power, the researchers employed two unsupervised machine learning methods—association analysis and cluster analysis—to determine the common attributes of students in the pass and fail groups. The researchers found 150 rules and six clusters (i.e., four of these clusters belong to the "pass" group and the other two belong to the "fail" group). Although the researchers began with two predefined groups (i.e., pass and fail), they still adopted a person-centered approach of uncovering latent patterns in these two groups by employing association rules analysis and cluster analysis. Such a situation showcases the complexity and coexistence of person- and variable-centeredness in a single research design. However, they did not stop there.

The latter part of the research design showcases elements of person-centeredness much in the way we observed with cluster analysis. In this case, the students in the potentially poor performance groups were interviewed to learn about their lived experiences, which were compared to the states identified by one of the models (i.e., a Hidden Markov Model). Like Faber and Benson's (2017) research design, this qualitative layer contextualizes the data-driven method using a person-centered approach, conducting a follow-up qualitative analysis to inform the interpretation of the quantitative model results.

EMBRACING AND EXPLORING OUTLIERS

When using traditional statistical techniques, the researcher must be mindful of data points too far from the average because such points can have non-trivial impacts on the results. Although some techniques, like cluster analysis, generally do not have many assumptions that must be satisfied before applying them to a dataset, ignoring observations believed to be “outliers” in the sample could still potentially cause issues with the modeling. For example, Nelson et al. (2015) employed cluster analysis to investigate the learning profiles adopted by students in entry-level courses, where self-regulation, motivation and affect, future time perspective, and course affect were used as input variables. During preprocessing, the authors excluded 10 percent of the data points before applying a two-step cluster analysis, for which the first step used sequential clustering and the second adopted agglomerative hierarchical clustering to avoid the potential effects of extreme values.

Typical advice about handling outliers leans toward a process of screening and removal (e.g., Osborne & Overbay 2004); however, other authors have illustrated the statistical consequences of removing outliers, such as André (2022), who argues that such procedures may inflate Type I errors. In the context of Nelson et al. (2015), 10 percent is a high proportion of the dataset to exclude without examining the underlying reasons for the sample distribution—the considerations of which were not provided. Accordingly, simply excluding these data points can potentially create bias and exclude the perspectives from smaller subgroups of participants. This approach of handling outliers by removing individuals from the sample without further examination showcases the variable-centeredness of seeking for homogeneity of the sample and potentially excluding individuals' voices from small subgroups, which relates to the QuantCrit principle, “numbers are not neutral.”

On the other hand, researchers can embrace outliers in their analyses—which could take several forms. Once again, Faber and Benson (2017) demonstrated one way to identify and handle “outliers” using a person-centered approach: leaning into a contradictory observation relative to the cluster analysis findings. When examining the intra-cluster similarity in one group during the qualitative analysis, a student was found to have no epistemic aim, which contrasted with the other interviewed members of the same cluster. In this case, the outlier was not based on numerical issues stemming from a stray data point; instead, the outlier emerged when qualitative evidence was juxtaposed with quantitative findings. Instead of dismissing this student as an outlier, the researchers carefully examined the student's responses to understand why the student differed from the rest. This approach manifests the person-centered element of preserving the diversity in individual's responses (Godwin et al. 2021) and embodies the QuantCrit principle that data cannot speak for themselves; it moves beyond a static description of the cluster's “average” and explores counteractive variation within algorithm identified group which the members are supposed to share common characteristics. In a mixed methods design, this person-centered approach is particularly robust for bolstering research quality and avoiding de-emphasizing potentially minoritized voices in the sample.

In Faber and Benson's (2017) case, they engaged in a practice similar to negative case analysis in qualitative research by digging deeper into the individuals' experience in the subgroup to understand why the divergence emerged. Negative case analysis involves modifying existing hypotheses by examining cases that go against them—actively seeking outliers (Kidder 1981; Newman & Benz 1998)—although this does not preclude authors from leaning into divergence (Creamer & Edwards 2019).

Not all data-driven algorithms are designed to assign individuals to a single group; some techniques enable researchers to implement less strict membership rules. For example, Scheidt et al. (2021) aimed to understand how engineering students can be grouped by their noncognitive and affective factors, which included the *big five* personality traits (i.e., neuroticism, extraversion, agreeableness, conscientiousness, and openness), grit, mindset, and mindfulness. Scheidt et al. (2021) recognized that it was dubious all data points could be perfectly assigned into a set number of clusters; moreover, they realized that most deterministic clustering algorithms would force each individual into a single cluster and not allow partial membership—even though some of the students might not be similar to other members in the cluster. To reflect the reality that not every student can fit into a single profile and to avoid the misassignment of students to clusters because all points are given a label, they chose to use a Gaussian Mixture Model (GMM). Unlike conventional clustering algorithms, GMM is a probabilistic clustering method that allows partial membership among clusters and un-clustered individuals.

Choosing such a clustering algorithm not only aligns with the premise of person-centered approaches by recognizing and preserving the heterogeneity in the sample, but it is also congruent with the QuantCrit principle, “categories are neither natural nor given.” Researchers aiming to perform a similar analysis could benefit significantly from probabilistic approaches to finding latent subgroups. Allowing for some uncertainty avoids ascribing too much weight to a given characterization of the individual. Instead, the individual can be seen as exhibiting qualities of different profiles.

TRACKING MOVEMENT BETWEEN CLUSTERS LONGITUDINALLY

The terminology distinguishing person-centered and variable-centered approaches can be traced back to Block (1971) in a longitudinal study of personality development. Naturally, it is fitting to see how engineering education researchers combine this origin of person-centered approaches with data-driven methods to explore how students transition between latent categories. Cakir and Gheorghe (2017) offer an example in which they adopted a two-step cluster analysis to investigate students’ academic profiles and identify the factors affecting students’ academic performance. For the first step of clustering, a hierarchical clustering algorithm was used to group courses and define the stages based on the course grouping and context, where three stages were identified for technical courses: stage one was composed of introductory coursework such as calculus, physics, and chemistry, stage two included earlier content in industrial engineering, such as statistics and linear algebra, and stage three referred to the junior and senior level coursework. The second step of their clustering process was to use an expectation maximization clustering algorithm to construct three general student academic profiles over four years, representing low, medium, and high performance. After the stages for the courses were defined and three academic profiles were constructed, students’ membership of the academic profile clusters at the different stages was tracked to unfold their academic performance trajectories. This application of cluster analysis provides a novel way of utilizing cluster analysis as a person-centered approach (i.e., personalizing the academic performance trajectory) by not only creating stages and constructing academic profiles but also showing the change in these latent categories over time.

EXPLORING SOCIAL DIMENSIONS OF RACE AS VARIABLES

By carefully selecting variables, one could take a person-centered approach to create more authentic groupings that avoid relying on demographic categories as proxies for broader constructs—most notably, racism. For example, Martin et al. (2015) employed cluster analysis to uncover the social capital profile (i.e., the characterization of students’ social networks and their access to social resources, such as resources provided by kin, friends, and education) of engineering students and investigated how race/ethnicity and gender affect students’ social capital profiles. In their preprocessing, they purposefully excluded students’ demographics as commonly used variables to form groups and constructed their social capital profile instead using variables like the size of their social network, the strength of those ties, the heterophily of those

connections, and what resources their connections provide them. This implementation of cluster analysis, in combination with the careful handling of race/ethnicity and other dichotomized social notations, such as gender, by Martin et al. (2015), saliently echoed the concern of the first principle of QuantCrit—the centrality of racism. They argued that using cluster analysis to find latent groupings based on students' social capital characteristics provides researchers with a more comprehensive lens to view complicated social capital problems compared to simply using pre-determined demographics as the sole categories.

Gillborn et al. (2018) underscored the socially constructed nature of race, and it is not often obvious how one might quantify what we discuss as racism or the impact of racism. Other researchers studying racialized and gendered topics can benefit from collecting these kinds of data to better capture the socialized dimensions of each individual's fixed characteristics.

EXPLORING THE BRANCHES OF A DECISION TREE FOR LATENT GROUPS

Shifting from the myriad examples of cluster analysis, this approach was specific to the data-driven method, *decision trees*. Decision trees, by design, work by recursively partitioning the input space into regions referred to as “leaves,” which gradually filter the input into a category based on a series of conditional statements (e.g., “mother attended college?” *no* → “father attended college?” *no* → “first-generation student.”).

Demographic categories are often used as inputs to inform the classification. For example, Inkelas et al. (2021) employed decision trees to build a model for predicting incoming college students' choices of their first calculus course. Incoming students' high school academic achievement, SAT scores, and demographic information such as gender and race/ethnicity were inspected as input features for the classifier. Even though gender and race/ethnicity were excluded from the final decision tree model because of their subpar predictive power, the purpose of building a predictive model using demographic variables showcases the variable-centeredness of adopting the classification tree method in this study. The intention to use predefined and often dichotomized social notations such as gender and race as candidate predictors highlights the concern of two QuantCrit principles, “centrality of racism” and “categories/groups are neither natural nor given.” Using socialized labels like race can lead one down a path of “racial reasoning” (Zuberi 2008, p. 131) and can obscure the causal mechanisms behind the classification (Gillborn et al. 2018).

Similarly, Tan et al. (2021) used a random forest model, an ensemble learning algorithm based on decision trees, to rank the predictors of engineering major choices of students by their predictive power. The predictors they examined included students' demographic categories, high school characteristics and experience, and standardized test scores. Students' gender was fed into the classifier and was recognized by the authors as the most important predictor of student engineering major choice. Tan et al. (2021) wrote that there is “critical predictive power in forecasting students declaring an engineering major” (p. 581) and that “removing the gender variable from the pool of explanatory variables means a 28% (Panel a [of variable importance plot]) and 19% (Panel b [of variable importance plot]) loss the total prediction accuracy” (p. 582). The authors recognize that the random forest method is a black box because of its opaqueness regarding interpretability, which, when combined with socialized identities, can muddy the understanding of precisely what is driving different phenomena, such as engineering major choice. Therefore, to avoid potential discrimination and increase the interpretability of the results, it is essential for researchers who adopt such “black box”-like machine learning algorithms to contextualize, investigate, and justify the selection of variables fed into the model, especially when they are related to individual's social identities or demographics.

However, person-centered applications of decision trees were present in the sampled articles, such as using the branches of the decision tree as a basis for forming latent groups. In the previous two examples (Inkelas et al. 2021; Tan et al. 2021), decision trees were implemented in a standard variable-centered way, but Singer et al. (2020) demonstrated how to deploy a decision tree classifier using both person- and variable-centered approaches. Singer et al. (2020) used decision trees to investigate the predictive power of certain variables, such as age, gender, and course

grades, to predict the academic stability of students with learning disabilities, as measured by the standard deviation in weighted average grades over eight terms; in particular, the researchers wanted to explore if including learning disabilities and accommodations as input variables can significantly affect such a model. Nonetheless, another reason drove the researchers to adopt the decision tree method as the classification method compared to other standard methods, such as neural networks. Unlike random forest models, the mechanism by which a single decision tree filters the data into categories is much more clearly laid out in its branches, and these branches can be used to define latent groupings based on subsets of splits in categories. The researchers contended that finding these subgroups of students can help instructors or professionals provide more effective accommodations based on students' identified profiles. Thus, even though the primary goal of this study carries a classic variable-centeredness, the researchers implemented the decision tree method for a person-centered purpose—identifying latent groupings—and this implementation embodies the QuantCrit principle “categories/groups are not natural nor given.”

Similar to Singer et al.'s (2020) study, Kaleita et al. (2016) used a decision tree to identify a set of variables that can be used to predict if students will be “at-risk” (i.e., GPA <2.0) before they start their first semester in college, such as ACT scores, demographic categories, and major certainty. The purpose of selecting predictors to build a model showcases the variable-centeredness of adopting decision tree methods—after all, decision trees are techniques designed to bin students into pre-labeled groups. However, Kaleita et al. (2016) argued that there might have been a situation in which a subgroup of students could be impacted heavily by some factors, but these factors do not affect others to the same extent. This consideration recognizes the latent diversity that could exist in the sample and led the authors to uncover these groupings to personalize the interventions for potential at-risk students.

DISCUSSION

In this discussion, we will highlight some cross-cutting ideas identified through the review to expand our results. In particular, we will discuss the prominence of cluster analysis and principal component analysis as data-driven methods in our sample to affirm the idea that methods are not inherently person-centered or variable-centered. We will also describe potential issues with operationalizing variables and data collection that could be the foundation for future work in this area.

CLUSTER ANALYSIS WAS THE MOST POPULAR METHOD AND MOST LIKELY TO BE USED IN DIVERSE PERSON-CENTERED APPLICATIONS

Cluster analysis was employed in more than half the reviewed articles ($n = 13$), which demonstrates its popularity as a data-driven method in mainstream engineering education research. Lund and Ma (2021) describe cluster analysis as a data mining algorithm usually employed to uncover latent patterns and group observations in a sample based on the input features; accordingly, it is well-suited for a person-centered approach. There are several different methodologies for clustering data, which include centroid/partition-based (e.g., k -means), connectivity-based (e.g., hierarchical), density-based (e.g., DBSCAN), and model-based (e.g., Gaussian Mixture Models), and fuzzy-based (e.g., c -means) (Tan et al. 2005). Although there is considerable diversity in the approaches researchers can take to group observations together, the centroid-based algorithm called k -means has grown in popularity in practice (Ashabi et al. 2020). Table 3 tabulates the different clustering algorithms used by researchers as the main quantitative method in their research; as noted in the table, the variety of algorithms used demonstrates the engineering education research community has experimented well beyond the popular k -means. Moreover, the diversity in the algorithms is rivaled by the multiple ways cluster analysis was used in a person-centered fashion by leveraging its flexibility in forming different latent subgroups in a sample.

ARTICLE	CLUSTERING ALGORITHM
Faber & Benson (2017)	<i>k</i> -means
Scheid et al. (2021)	Gaussian Mixture Model
Marbouti et al. (2021)	Bisecting <i>k</i> -means
Jaiswal et al. (2021)	Hierarchical clustering
Martin et al. (2015)	Two-step clustering (not specified)
Yellamraju et al. (2019)	<i>n</i> -TARP clustering
Ruipérez-Valiente et al. (2017)	Two-step clustering (not specified)
Reid et al. (2016)	McDermott's three-stage cluster analysis
Cakir & Gheorghe (2017)	Two-step clustering (hierarchical → expectation maximization)
S. Haase (2014)	Nearest centroid
Nelson et al. (2015)	Two-step clustering (not specified)
Gallego et al. (2016)	Hierarchical clustering
Choe & Borrego (2020)	Two-step clustering (<i>k</i> -means variant → modified hierarchical)

Table 3 Summary of cluster analysis applications in sample (*n* = 13).

PRINCIPAL COMPONENT ANALYSIS DOES NOT SEEM COMPATIBLE WITH PERSON-CENTERED APPROACHES

PCA, a commonly employed technique for dimension reduction, seeks to identify a set of composite variables that can effectively represent a dataset in a lower dimension while preserving essential information (Kherif & Latypova 2020). We found that authors typically employed principal component analysis with a variable-centric perspective. This is consistent with Godwin et al.'s (2021) initial assessment of PCA as a technique most suited to variable-centered aims. In our sample, PCA was only used twice as the primary technique in the research design but appeared frequently alongside other data-driven techniques. Within these applications, using PCA did not seem compatible with a person-centered approach.

To elaborate on why, consider the process of choosing variables. Besides deciding whom to include and exclude, what to measure is another common analytical decision the researchers need to make before implementing data-driven techniques. Such variable selection processes can be performed using a variable-centered approach in cases where the dimensionality of the data is too large. Because PCA is a dimensionality reduction technique, the studies in the sample would use it to preprocess data to reduce the number of variables at play. For example, Martin and Sorhaindo (2019) examined intrinsic and extrinsic motivational factors as predictors of academic achievement in civil engineering students. They used PCA to consolidate the motivational factors into distinct intrinsic and extrinsic groups. The initial set of 22 motivation variables was reduced to five principal components, explaining 66% of the variance. This application of PCA to combine variables highlights aspects of variable-centeredness, which embodies the QuantCrit principle, “numbers are not neutral.” The researchers replaced the original set of variables with the derived principal components—even though a reasonably large variation loss (i.e., 34%) was recognized. After conducting the PCA, the researchers probed for differences between groups defined by their demographic characteristics, such as gender (male vs. female), national origin (native vs. international), and age.

There are two points to raise here. First, considerable diversity is washed out during aggregation and can bring about misleading conclusions if taken at face value (Shafer et al. 2021). By combining the subgroups comprising the native and international, an aggregation fallacy occurs similar to the underrepresented and minority category often seen in studies with low *n*'s in populations with racial identities like Black, Latine, and Native American (Xu et al. 2023). These explorations can work against the QuantCrit principle of “categories/groups are neither natural nor given,” particularly in the case of national origin, a dichotomized and highly aggregated category in this study.

Second, principal components are not necessarily informed by theory nor guaranteed to be interpretable and theoretically meaningful. By comparing demographic groups to one another using these uniquely constructed variables, the interpretations of statistical tests can be misused or lead future work down fruitless paths because of spurious correlations in variables from a foundational study. Pairing principal component analysis with a person-centered approach generally does not offset these issues. For example, Gallego et al. (2016) investigated how workload affects students' academic performance by adopting cluster analysis to group students with similar characteristics by examining the efficiency rate (ER, which is the number of passing students/the number of enrolled students), success rate (SR which is the number of passing students/the number of attending students), drop-out rate (DR which is the number of non-attending students/the number of enrolled students), time (student's rating of time needed to pass the course), and difficulty (the student's perceived level of difficulty of a subject).

Before applying cluster analysis, they used principal component analysis (PCA) on the entire set of candidate variables, then chose the two resulting principal components as the aggregated variables to be used to form clusters. These principal components accounted for 70% of the total variance in the sample. Although PCA can aid the researcher by reducing the total number of variables to a more workable number that can be visualized or better aligned with a desirable variable-to-case ratio, the resulting principal components are not guaranteed to be interpretable or theoretically meaningful. In this case, the first principal component was primarily a combination of three variables: ER, SR, and Difficulty. The researchers proposed that this principal component could be interpreted as the "number of students passing the subject" (Gallego et al. 2016, p. 9) because ER and SR are related to the measures of academic performance. However, the other highly associated variable, "Difficulty," was not clearly addressed, which muddies the interpretation of the composite variable.

Without a clear operationalization, the resulting clusters are likely to be less meaningful, and drawing conclusions from these principal components for classroom interventions can lead the researchers down unproductive paths. The approach to preprocessing, that is, employing PCA to form composite variables, aligns with a variable-centered approach—considering this step explicitly aims to find homogeneity among variables to reduce the data's dimensionality. Moreover, using principal components without thoroughly interrogating their meaning works against the QuantCrit principle, "numbers are not neutral." These composite variables are highly contextualized with respect to the data—meaning the actual composite variables from dataset to dataset with the same variables are not guaranteed to be the same. The potential for spurious correlations in high-dimensional datasets is not trivial, so further work based on these principal components could lead to misuse if care is not taken when operationalizing such variables.

(DATA-DRIVEN) METHODS ARE NOT INHERENTLY PERSON-CENTERED OR VARIABLE-CENTERED

At this point, we contrasted cluster analysis and principal component analysis in terms of person-centered and variable-centered approaches. Although it might be tempting to deem a particular method person-centered or variable-centered, it became clear throughout this review that this thinking can be limiting. For example, cluster analysis is perhaps the most likely technique for researchers to use in a person-centered fashion. In fact, Godwin et al. (2021) associate cluster analysis's ability to identify latent data structures with person-centered analyses. This aligns with the premise of using a latent diversity perspective when conducting a quantitative analysis to understand relationships among individuals in a sample beyond demographic categories (see Godwin 2017). Although the goal of cluster analysis can be stated simply, the way the technique was used within our sample is a testament to the myriad decisions involved with its implementation.

Although cluster analysis is an inductive technique by design, in the sense that the groupings do not necessarily need to be known *a priori*, this does not preclude the researcher from applying the technique deductively in a variable-centric fashion. One example in the sample of such an approach was Reid et al. (2016). In their study, cluster analysis was utilized as a method

to justify the stability and repeatability of a psychometric instrument (i.e., Student Attitudinal Success Instrument or SASI-I), which is constructed to gather incoming engineering students' affective characteristics, such as intrinsic motivation, academic self-efficacy, and leadership. Using the factor analysis results of the SASI-I, the same number of clusters for each cohort from three years (i.e., student cohorts admitted in three different years) were generated, and the homogeneity coefficients for these clusters were found to be high between corresponding clusters over those years. These results were used as evidence for the stability of their instrument. Although subgroups of students were identified, as typical with applications of cluster analysis in these circumstances, the number of the clusters and the similarity of those clusters over years was used to support claims of validity evidence rather than describe the characteristics of the identified subgroups.

Similarly, Quan et al. (2019) selected five representative indicators and conducted a cluster analysis using k-means. They found that the two resulting clusters effectively grouped “at-risk” students (i.e., students who failed at least one course in a single semester), wherein one cluster had a high proportion of “at-risk” students, and the other had a high proportion of “not-at-risk” students. Thus, the cluster analysis result demonstrated the selected indicators' effectiveness – the two clusters are not only significantly different regarding selected features but also well separate at-risk students and not-at-risk students. Like Reid et al.'s (2016) use of cluster analysis to verify the stability of their instrument, this focus on using groupings of students as the medium to justify an instrument's usability is inherent to a variable-centric approach.

The other technique that seemed to be an entirely variable-centered method was the decision tree (and its associated approaches, like random forest). Godwin et al. (2021) binned decision trees squarely in the variable-centered category, explaining that these data-driven methods are focused on “reconcil[ing] variables with predefined (and thus potentially biased or inaccurate) categories” (p. 21). There is little to argue about this point because a decision tree is not necessarily designed to assign inputs to categories that are not explicitly used when developing the tree. Moreover, if decision trees are used to filter individuals into categories without careful consideration of the implications of those classifications, such as classifying students as *at-risk* using demographics among other variables (c.f., Brown 2016), there could be a potential for misuse. However, like how cluster analysis could be leveraged in a variable-centric fashion, examples of how researchers extracted some person-centered applications of decision trees in their studies were present in the sample, such as examining specific branches in the tree and treating the intermediate splits as latent groups. This thinking can generally be applied across quantitative techniques.

PERSON-CENTERED APPROACHES CANNOT COMPENSATE FOR POOR DATA COLLECTION OR OPERATIONALIZATION

Although we have discussed person-centered approaches in a positive light so far in this paper, it should be stressed that they are not a panacea for methodological issues in engineering education research. In particular, throughout the papers we discussed, and what is embedded in the usual premise of person-centered approaches, much of the focus is on using psychological measures (e.g., motivation, personality, epistemic beliefs) to group observations and draw comparisons of the latent diversity within the sample. For person-centered approaches to produce legitimate results at a fundamental level, the psychological measures under study must be appropriate in terms of their validity and reliability—or else the result will be a situation of *garbage in, garbage out*. More broadly, quantification is a key concern of QuantCrit throughout its principles. QuantCrit challenges the idea that concepts like racism are readily quantifiable and that numbers are inherently neutral (Gillborn et al. 2018). Using person-centered approaches does not address issues related to bias in data collection or operationalization of variables, and engineering education researchers should not see person-centered approaches as a fix for such issues.

As suggested by Castillo and Gillborn (2023), researchers should think critically about the metrics they use in their studies by reflecting on the origin of different measurements. Many psychological instruments have only been tested in specific contexts and are laden with cultural

and socioeconomic considerations; moreover, the psychological construct may have been operationalized in a way that is not congruent with the researcher's context. After all, validity is a continuous process, and selecting an instrument from a publication does not imply it will work elsewhere (Douglas & Purzer, 2015). These points underscore that much of the success of a person-centered approach is reliant on the researcher engaging critically with design decisions at the onset of the study. Although bias is never truly eliminated from the analyses, pinpointing sources of bias in their data collection and operationalizations initially better position engineering education researchers to frame the limitations of their work when the time comes to report on their findings.

CONCLUSION

This review focused on identifying the person-centered and variable-centered ways in which engineering education researchers use data-driven methods. The review sampled from four popular engineering education journals over the course of a decade and revealed the popularity of cluster analysis in performing a wide variety of tasks. These included both person-centered purposes, like finding latent subgroups at one timepoint and longitudinally, and variable-centered purposes, like generating evidence of an instrument's stability over time. Similarly, we observed person-centered ways to implement a decision tree model that can go beyond its more obvious application as a classifier, which is variable-centric.

Godwin et al. (2021) explained that implementing classification and regression trees tends to be variable-centric because such methods align variables with pre-determined categories. However, this interpretation primarily focuses on the broad goals of a decision tree approach. When we zoom out and examine what can be done with a decision tree once constructed, person-centered approaches become more apparent—as we saw in the sample. Singer et al. (2020) and Kaleita et al. (2016) demonstrated the possibility of implementing decision trees in a person-centered fashion by exploring latent subgroups that emerge if one traces through the branches of the decision tree. Using the decision tree, a data-driven method with some intrinsic variable-centered attributes by design, in a person-centered way, shows that researchers can balance the potential drawbacks brought by variable-centered analysis and deconstruct how the model emerged from the data. These post-hoc explorations can invite subsequent analyses that build off the patterns. Much like our findings from papers in the sample using cluster analysis, decision trees are pre-disposed to being used in specific ways that align with variable-centered principles. However, this does not preclude them from being used in a person-centered analysis.

FUTURE WORK

One direction for future work can be increasing the timespan of consideration and publication venue, which has the potential to enrich the findings demonstrated in this paper and achieve a more comprehensive and overarching SLR. One large database of publications is the paper repository hosted by the American Society for Engineering Education, which contains conference proceedings for its annual conference and affiliated conferences like the First Year Engineering Experience conference series. Another popular conference that would be useful to review papers from is the Frontiers in Education conference hosted by the Institute for Electrical and Electronic Engineers (IEEE). More data-driven methods, diverse applications, and evolutions of such approaches might emerge.

Another direction for future research can involve focusing on applying a specific data-driven method, such as cluster analysis or decision trees, in engineering education research. We have showcased how versatile some data-driven methods can be when they are applied in different ways—person-centered, variable-centered, or both—according to different research purposes and analytical decisions made by researchers. Thus, such method-specific research is warranted, which can provide a more in-depth understanding of the current state of specific data-driven methods in engineering education research and offer researchers informative exemplars on when and how to apply the data-driven method creatively and effectively.

Finally, although not the focus of this paper, another line of inquiry that can sprout from this work is examining what variables researchers are using to conduct person-centered approaches. As we noted in our discussion, many of the variables considered were psychological measurements. The degree to which these measurements and the conclusions drawn from them are interrogated for validity and reliability—including how the authors reflect with criticality through a QuantCrit lens—can be performed in a deeper analysis of engineering education publications.

LIMITATIONS

As mentioned in the “Future work” section of the Conclusion, the main limitation of the current work is the timespan of the reviewed literature. This SLR only considered publications for empirical, human subjects research from 2011 to 2021 in four leading engineering education journals—which yielded a specific subset of data-driven method applications. If any of the constraints were relaxed on the inclusion criteria, the results could likely appear considerably different. Thus, this work only depicts a detailed picture of data-driven methods through the lens of variable- and person-centeredness with a critical eye on a small portion of existing literature. The results should not be taken as a comprehensive review of how data-driven methods are applied in engineering education for broader settings (e.g., research on non-human subjects and looking back from 2010).

A second limitation is our scope in the search string. When forming our search string, we included typical keywords like “machine learning” and “data mining” among more specific techniques commonly associated with data-driven methods. Considering the search string was informed by mainstream data-driven methods, it likely does not capture the full range of niche techniques that might have been employed in the sampled journals. Therefore, a selection of papers could have been overlooked. Moreover, depending on how one defines a data-driven or machine learning method, the boundaries of what would ultimately be included in the review would be reshaped. Thus, the results should not be interpreted as being a complete treatment of all possible data-driven methods.

It is anticipated that this work can encourage more comprehensive studies about the deployment of data-driven methods in engineering education research. Future work can provide researchers with different perspectives to assist them with finding novel ways to balance the benefits and biases brought when adopting powerful data-driven methods in their research.

FINAL THOUGHTS

As described in the delineation between person-centered and variable-centered approaches, these must coexist and not be viewed as a dichotomy. In the particular case of data-driven approaches, which are aligned with research goals that are often person-centric in nature, we have seen several tangible examples that engineering education researchers can use to provide a balance in their design that can zoom both in and out on the latent characteristics of their sample. In fact, several of them were rooted in situating the data-driven methods in a mixed methods design, which intersected the technique with qualitative approaches to contextualize the findings. However, as we demonstrated, researchers need not always use a mixed methods design to humanize their approach. In particular, by embracing a latent diversity perspective in their research design (Godwin 2017), researchers can avoid an essentialized perspective on the role of race (and other sensitive constructs) in their work. We anticipate spotlighting these strategies will provide engineering education researchers with an expanded, practical toolkit for engaging in a QuantCrit mindset while handling tools that can cause considerable harm if misused.

APPENDIX A. REVIEWED PAPERS

- Cakir, V., & Gheorghe, A.** (2017). Longitudinal academic performance analysis using a two-step clustering methodology. *International Journal of Engineering Education*, 33(1 A), 203–215.
- Choe, N. H., & Borrego, M.** (2020). Master's and doctoral engineering students' interest in industry, academia, and government careers. *Journal of Engineering Education*, 109(2), 325–346. <https://doi.org/10.1002/jee.20317>

- Cutumisu, M., & Guo, Q. (2019). Using topic modeling to extract pre-service teachers' understandings of computational thinking from their coding reflections. *IEEE Transactions on Education*, 62(4), 325–332. <https://doi.org/10.1109/TE.2019.2925253>
- Faber, C., & Benson, L. C. (2017). Engineering students' epistemic cognition in the context of problem solving. *Journal of Engineering Education*, 106(4), 677–709. <https://doi.org/10.1002/jee.20183>
- Finj, E. H., Awadallah, F., Parast, M. M., & Abu-Lebdeh, T. (2018). The impact of project-based learning on improving student learning outcomes of sustainability concepts in transportation engineering courses. *European Journal of Engineering Education*, 43(3), 473–488. <https://doi.org/10.1080/03043797.2017.1393045>
- Gallego, E., Diaz Barcos, V., Correa Hernando, E. C., Sanchez Espinosa, E., & Callejo Ramos, A. (2016). Influence of the perceived workload of students on the academic performance rates. *International Journal of Engineering Education*, 32(2), 670–681. <https://doi.org/10.31578/jebs.v6i1.218>
- Haase, S. (2014). Engineering students' sustainability approaches. *European Journal of Engineering Education*, 39(3), 247–271. <https://doi.org/10.1080/03043797.2013.858103>
- Inkelas, K. K., Maeng, J. L., Williams, A. L., & Jones, J. S. (2021). Another form of undermatching? A mixed-methods examination of first-year engineering students' calculus placement. *Journal of Engineering Education*, 110(3), 594–615. <https://doi.org/10.1002/jee.20406>
- Jaiswal, A., Lyon, J. A., Zhang, Y., & Magana, A. J. (2021). Supporting student reflective practices through modelling-based learning assignments. *European Journal of Engineering Education*, 46(6), 987–1006. <https://doi.org/10.1080/03043797.2021.1952164>
- Kaleita, A. L., Forbes, G. R., Ralston, E., Compton, J. I., Wohlgemuth, D., & Raman, D. R. (2016). Pre-enrollment identification of at-risk students in a large engineering college. *International Journal of Engineering Education*, 32(4), 1647–1659.
- Ko, C.-Y., & Leu, F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education*, 64(1), 50–57. <https://doi.org/10.1109/TE.2020.3004596>
- Kolmos, A., Mejlgaard, N., Haase, S., & Holgaard, J. E. (2013). Motivational factors, gender and engineering education. *European Journal of Engineering Education*, 38(3), 340–358. <https://doi.org/10.1080/03043797.2013.794198>
- Marbouti, F., Ulas, J., & Wang, C.-H. (2021). Academic and demographic cluster analysis of engineering student success. *IEEE Transactions on Education*, 64(3), 261–266. <https://doi.org/10.1109/TE.2020.3036824>
- Martín, H., & Sorhaindo, C. (2019). A comparison of intrinsic and extrinsic motivational factors as predictors of civil engineering students' academic success. *International Journal of Engineering Education*, 35(2), 458–472.
- Martin, J. P., Brown, S., Miller, M. K., & Steff, S. K. (2015). Characterizing engineering student social capital in relation to demographics. *International Journal of Engineering Education*, 31(4), 914–926.
- Mou, C., Zhou, Q., & Zou, X. (2017). Understanding and predicting poor performance of computer science students from short time series test results. *International Journal of Engineering Education*, 33(6), 1803–1814.
- Nelson, K. G., Shell, D. F., Husman, J., Fishman, E. J., & Soh, L.-K. (2015). Motivational and self-regulated learning profiles of students taking a foundational engineering course: Learning profiles of students in a foundational engineering course. *Journal of Engineering Education*, 104(1), 74–100. <https://doi.org/10.1002/jee.20066>
- Quan, W., Zhou, Q., Zhong, Y., & Wang, P. (2019). Predicting at-risk students using campus meal consumption records. *International Journal of Engineering Education*, 35(2), 563–571.
- Reid, K., Imbrie, P. K., Lin, J. J., Reed, T., & Immekus, J. C. (2016). Psychometric properties and stability of the student attitudinal success instrument: the SASI-I. *International Journal of Engineering Education*, 32(6), 2470–2486.
- Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). Detecting and clustering students by their gamification behavior with badges: A case study in engineering education. *International Journal of Engineering Education*, 33(2-B), 816–830.
- Scheidt, M., Godwin, A., Berger, E., Chen, J., Self, B. P., Widmann, J. M., & Gates, A. Q. (2021). Engineering students' noncognitive and affective factors: Group differences from cluster analysis. *Journal of Engineering Education*, 110(2), 343–370. <https://doi.org/10.1002/jee.20386>
- Singer, G., Golan, M., Rabin, N., & Kleper, D. (2020). Evaluation of the effect of learning disabilities and accommodations on the prediction of the stability of academic behaviour of undergraduate engineering students using decision trees. *European Journal of Engineering Education*, 45(4), 614–630. <https://doi.org/10.1080/03043797.2019.1677560>

- Tadayon, M., & Pottie, G. J.** (2020). Predicting student performance in an educational game using a hidden markov model. *IEEE Transactions on Education*, 63(4), 299–304. <https://doi.org/10.1109/TE.2020.2984900>
- Tan, L., Main, J. B., & Darolia, R.** (2021). Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. *Journal of Engineering Education*, 110(3), 572–593. <https://doi.org/10.1002/jee.20393>
- Yellamraju, T., Magana, A. J., & Boutin, M.** (2019). Investigating students' habits of mind in a course on digital signal processing. *IEEE Transactions on Education*, 62(4), 312–324. <https://doi.org/10.1109/TE.2019.2924610>

DATA ACCESSIBILITY STATEMENT

Appendix A contains the list of articles used in this review for the purpose of reproducibility.

ACKNOWLEDGEMENTS

This paper is based on a thesis submitted by Jiafu Niu at the University of Cincinnati in partial fulfillment of an MS degree in Engineering Education (Niu 2023). Moreover, portions of this paper were previously reported in a work-in-progress paper published in the American Society for Engineering Education proceedings for the 2023 Annual Conference and Exposition (Niu & Reeping 2023).

FUNDING INFORMATION

This research was supported by a grant from the American Educational Research Association – Division D (Measurement & Research Methodologies).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Jiafu Niu conducted the study, including data collection and analysis, under the supervision of David Reeping. David provided guidance by reviewing the results and collaboratively synthesizing the findings with Jiafu. Both authors contributed to the manuscript's writing.

AUTHOR AFFILIATIONS

Jiafu Niu

University of Cincinnati, US

David Reeping  orcid.org/0000-0002-0803-7532

University of Cincinnati, US

REFERENCES

- Anderson, C.** (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 16(7), 16–07.
- André, Q.** (2022). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, 151(1), 213–223. <https://doi.org/10.1037/xge0001069>
- Ashabi, A., Sahibuddin, S. B., & Salkhordeh Haghighi, M.** (2020, December). The systematic review of K-means clustering algorithm. In *2020 The 9th International Conference on Networks, Communication and Computing* (pp. 13–18). <https://doi.org/10.1145/3447654.3447657>
- Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E.** (2019). *Public knowledge and experiences with data-driven ads*. Pew Research Center. <https://www.pewresearch.org/internet/2019/11/15/public-knowledge-and-experiences-with-data-driven-ads/>

- Baker, R. S., & Yacef, K.** (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- Block, J.** (1971). *Lives through time*. Bancroft Books.
- Borrego, M., Foster, M. J., & Froyd, J. E.** (2014). Systematic literature reviews in engineering education and other developing interdisciplinary fields: Systematic literature reviews in engineering education. *Journal of Engineering Education*, 103(1), 45–76. <https://doi.org/10.1002/jee.20038>
- Brown, K.** (2016). *After the “at-risk” label: Reorienting educational policy and practice*. Teachers College Press.
- Cakir, V., & Gheorghe, A.** (2017). Longitudinal academic performance analysis using a two-step clustering methodology. *International Journal of Engineering Education*, 33(1 A), 203–215.
- Castillo, W., & Gillborn, D.** (2023). How to “QuantCrit:” Practices and questions for education data researchers and users. (EdWorkingPaper: 22–546). Annenberg Institute at Brown University. <https://doi.org/10.26300/v5kh-dd65>
- Chan, C. K. Y., & Fong, E. T. Y.** (2018). Disciplinary differences and implications for the development of generic skills: A study of engineering and business students’ perceptions of generic skills. *European Journal of Engineering Education*, 43(6), 927–949. <https://doi.org/10.1080/03043797.2018.1462766>
- Choe, N. H., & Borrego, M.** (2020). Master’s and doctoral engineering students’ interest in industry, academia, and government careers. *Journal of Engineering Education*, 109(2), 325–346. <https://doi.org/10.1002/jee.20317>
- Creamer, E., & Edwards, C.** (2019). Embedding the dialogic in mixed method approaches to theory development. *International Journal of Research & Method in Education*, 42(3), 239–251. <https://doi.org/10.1080/1743727X.2019.1598357>
- Cunningham-Nelson, S., Baktashmotlagh, M., & Boles, W.** (2019). Visualizing student opinion through text analysis. *IEEE Transactions on Education*, 62(4), 305–311. <https://doi.org/10.1109/TE.2019.2924385>
- Cutumisu, M., & Guo, Q.** (2019). Using topic modeling to extract pre-service teachers’ understandings of computational thinking from their coding reflections. *IEEE Transactions on Education*, 62(4), 325–332. <https://doi.org/10.1109/TE.2019.2925253>
- Douglas, K. A., & Purzer, S.** (2015). Validity: Meaning and relevancy in assessment for engineering education research. *Journal of Engineering Education*, 104(2), 108–118. <https://doi.org/10.1002/jee.20070>
- Faber, C., & Benson, L. C.** (2017). Engineering students’ epistemic cognition in the context of problem solving. *Journal of Engineering Education*, 106(4), 677–709. <https://doi.org/10.1002/jee.20183>
- Gallego, E., Diaz Barcos, V., Correa Hernando, E. C., Sanchez Espinosa, E., & Callejo Ramos, A.** (2016). Influence of the perceived workload of students on the academic performance rates. *International Journal of Engineering Education*, 32(2), 670–681.
- Gillborn, D., Warmington, P., & Demack, S.** (2018). QuantCrit: Education, policy, ‘Big Data’ and principles for a critical race theory of statistics. *Race Ethnicity and Education*, 21(2), 158–179. <https://doi.org/10.1080/13613324.2017.1377417>
- Godwin, A.** (2017). *Unpacking latent diversity*. Paper presented at the ASEE Annual Conference and Exposition, Columbus, OH. <https://peer.asee.org/unpacking-latent-diversity>. <https://doi.org/10.18260/1-2--29062>
- Godwin, A., Benedict, B., Rohde, J., Thielmeyer, A., Perkins, H., Major, J., Clements, H., & Chen, Z.** (2021). New epistemological perspectives on quantitative methods: An example using topological data analysis. *Studies in Engineering Education*, 2(1), 16. <https://doi.org/10.21061/see.18>
- Greene, J. C., Caracelli, V. J., & Graham, W. F.** (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11(3), 255–274. <https://doi.org/10.3102/01623737011003255>
- Haase, S.** (2014). Engineering students’ sustainability approaches. *European Journal of Engineering Education*, 39(3), 247–271. <https://doi.org/10.1080/03043797.2013.858103>
- Hampton, C., & Reeping, D.** (2019). Positionality: The Stories of Self that Impact Others. *2019 ASEE Annual Conference & Exposition Proceedings*, 33177. <https://doi.org/10.18260/1-2--33177>
- Hampton, C., Reeping, D., & Ozkan, D. S.** (2021). Positionality statements in engineering education research: A look at the hand that guides the methodological tools. *Studies in Engineering Education*, 1(2), Article 2. <https://doi.org/10.21061/see.13>
- Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., ... & Stachl, C.** (2021). Machine learning for the educational sciences. *Review of Education*, 9(3), e3310. <https://doi.org/10.1002/rev3.3310>
- Hofmans, J., Wille, B., & Schreurs, B.** (2020). Person-centered methods in vocational research. *Journal of Vocational Behavior*, 118(2020), 1–15. <https://doi.org/10.1016/j.jvb.2020.103398>
- Howard, M. C., & Hoffman, M. E.** (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods*, 21(4), 846–876. <https://doi.org/10.1177/1094428117744021>

- Inkelas, K. K., Maeng, J. L., Williams, A. L., & Jones, J. S. (2021). Another form of undermatching? A mixed-methods examination of first-year engineering students' calculus placement. *Journal of Engineering Education*, 110(3), 594–615. <https://doi.org/10.1002/jee.20406>
- Irving, G., & Askeff, A. (2019). AI safety needs social scientists. *Distill*, 4(2), e14. <https://doi.org/10.23915/distill.00014>
- Jaiswal, A., Lyon, J. A., Zhang, Y., & Magana, A. J. (2021). Supporting student reflective practices through modelling-based learning assignments. *European Journal of Engineering Education*, 46(6), 987–1006. <https://doi.org/10.1080/03043797.2021.1952164>
- Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 306–316). <https://doi.org/10.1145/3351095.3372829>
- Johri, A., Wang, G. A., Liu, X., & Madhavan, K. (2011, October). Utilizing topic modeling techniques to identify the emergence and growth of research topics in engineering education. In *2011 Frontiers in Education Conference (FIE)* (pp. T2F-1). IEEE. <https://doi.org/10.1109/FIE.2011.6142770>
- Kaleita, A. L., Forbes, G. R., Ralston, E., Compton, J. I., Wohlgemuth, D., & Raman, D. R. (2016). Pre-enrollment identification of at-risk students in a large engineering college. *International Journal of Engineering Education*, 32(4), 1647–1659.
- Kherif, F., & Latypova, A. (2020). Principal component analysis. In *Machine Learning* (pp. 209–225). Elsevier. <https://doi.org/10.1016/B978-0-12-815739-8.00012-2>
- Kidder, L. H. (1981). Qualitative research and quasi-experimental frameworks. In M. B. Brewer & B. E. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 226–256). Jossey-Bass.
- Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., & Silva, R. (2020, August). The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence* (pp. 616–626). PMLR. <https://arxiv.org/abs/1907.01040>
- Ko, C.-Y., & Leu, F.-Y. (2021). Examining successful attributes for undergraduate students by applying machine learning techniques. *IEEE Transactions on Education*, 64(1), 50–57. <https://doi.org/10.1109/TE.2020.3004596>
- Korkmaz, C., & Correia, A. P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, 56(3), 250–267. <https://doi.org/10.1080/09523987.2019.1669875>
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15), 5802–5805. <https://doi.org/10.1073/pnas.1218772110>
- Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37, 331–344. <https://doi.org/10.1007/s10462-011-9234-x>
- Kusner, M. J., & Loftus, J. R. (2020). *The long road to fairer algorithms*. Nature Publishing Group. <https://doi.org/10.1038/d41586-020-00274-3>
- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 1–18. <https://arxiv.org/abs/1703.06856>
- Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. *Handbook of Psychology*, 2, 691–716. <https://doi.org/10.1002/9781118133880.hop202024>
- Laursen, B. P., & Hoff, E. (2006). Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly*, 52(3), 377–389. <https://doi.org/10.1353/mpq.2006.0029>
- Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2018). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12), 2346–2363. <https://doi.org/10.1109/TKDE.2018.2876857>
- Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science research: K-means and k-medoids clustering. *Performance Measurement and Metrics*, 22(3), 161–173. <https://doi.org/10.1108/PMM-05-2021-0026>
- Marbouti, F., Ulas, J., & Wang, C.-H. (2021). Academic and demographic cluster analysis of engineering student success. *IEEE Transactions on Education*, 64(3), 261–266. <https://doi.org/10.1109/TE.2020.3036824>
- Martin, J. P., Brown, S., Miller, M. K., & Steff, S. K. (2015). Characterizing engineering student social capital in relation to demographics. *International Journal of Engineering Education*, 31(4), 914–926.
- Martín, H., & Sorhaindo, C. (2019). A comparison of intrinsic and extrinsic motivational factors as predictors of civil engineering students' academic success. *International Journal of Engineering Education*, 35(2), 458–472.
- McCutcheon, A. L. (1987). *Latent class analysis* (Vol. 64). Sage. <https://doi.org/10.4135/9781412984713>

- Morin, A. J. S., Gagne, M., & Bujacz, A. (2016). Feature topic: Person-centered methodologies in the organizational sciences. *Organizational Research Methods*, 19(1), 8–9. <http://journals.sagepub.com/doi/10.1177/1094428115617592>
- Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24(6), 882–891. <https://doi.org/10.1111/j.1530-0277.2000.tb02070.x>
- Nelson, K. G., Shell, D. F., Husman, J., Fishman, E. J., & Soh, L.-K. (2015). Motivational and self-regulated learning profiles of students taking a foundational engineering course: Learning profiles of students in a foundational engineering course. *Journal of Engineering Education*, 104(1), 74–100. <https://doi.org/10.1002/jee.20066>
- Newman, I., & Benz, C. R. (1998). *Qualitative-quantitative research methodology: Exploring the interactive continuum*. SIU Press.
- Niu, J. (2023). *Humanizing data-driven methods in engineering education research: A systematic literature review of four journals from 2011 to 2021* [Masters Thesis, University of Cincinnati]. ProQuest Dissertations & Theses Global. http://rave.ohiolink.edu/etdc/view?acc_num=ucin1704205635164482
- Niu, J., & Reeping, D. (2023). Work in progress: A systematic literature review of person-centered approaches and data-driven methods in engineering education research. *2023 ASEE Annual Conference & Exposition*. <https://doi.org/10.18260/1-2--44164>
- O’Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation*, 9(1), 6. <https://doi.org/10.7275/QF69-7K43>
- Pawley, A. L. (2017). Shifting the “default”: The case for making diversity the expected condition for engineering education and making whiteness and maleness visible. *Journal of Engineering Education*, 106(4), 531–533. <https://doi.org/10.1002/jee.20181>
- Pizard, S., & Vallespir, D. (2017). Towards a controlled vocabulary on software engineering education. *European Journal of Engineering Education*, 42(6), 927–943. <https://doi.org/10.1080/03043797.2016.1235139>
- Pluye, P., Grad, R. M., Levine, A., & Nicolau, B. (2009). Understanding divergence of quantitative and qualitative data (or results) in mixed methods studies. *International Journal of Multiple Research Approaches*, 3(1), 58–72. <https://doi.org/10.5172/mra.455.3.1.58>
- Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science*, 1(1), 59–66. <https://doi.org/10.1007/s42001-017-0013-6>
- Quan, W., Zhou, Q., Zhong, Y., & Wang, P. (2019). Predicting at-risk students using campus meal consumption records. *International Journal of Engineering Education*, 35(2), 563–571.
- Reeping, D., Lee, W., & London, J. (2023). Person-centered analyses in quantitative studies about broadening participation for Black engineering and computer science students. *Journal of Engineering Education*, 112(3), 769–795. <https://doi.org/10.1002/jee.20530>
- Reid, K., Imbrie, P. K., Lin, J. J., Reed, T., & Immekus, J. C. (2016). Psychometric properties and stability of the student attitudinal success instrument: the SASI-I. *International Journal of Engineering Education*, 32(6), 2470–2486.
- Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., & Delgado Kloos, C. (2017). Detecting and clustering students by their gamification behavior with badges: A case study in engineering education. *International Journal of Engineering Education*, 33(2-B), 816–830.
- Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed). Sage.
- Scheidt, M., Godwin, A., Berger, E., Chen, J., Self, B. P., Widmann, J. M., & Gates, A. Q. (2021). Engineering students’ noncognitive and affective factors: Group differences from cluster analysis. *Journal of Engineering Education*, 110(2), 343–370. <https://doi.org/10.1002/jee.20386>
- Shafer, D., Mahmood, M. S., & Stelzer, T. (2021). Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students. *Physical Review Physics Education Research*, 17(1), 1–13. <https://doi.org/10.1103/PhysRevPhysEducRes.17.010113>
- Shokri, R., Strobel, M., & Zick, Y. (2021). On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES ‘21)* (pp. 231–241). Association for Computing Machinery. <https://doi.org/10.1145/3461702.3462533>
- Singer, G., Golan, M., Rabin, N., & Kleper, D. (2020). Evaluation of the effect of learning disabilities and accommodations on the prediction of the stability of academic behaviour of undergraduate engineering students using decision trees. *European Journal of Engineering Education*, 45(4), 614–630. <https://doi.org/10.1080/03043797.2019.1677560>

- Tan, L., Main, J. B., & Darolia, R.** (2021). Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. *Journal of Engineering Education*, 110(3), 572–593. <https://doi.org/10.1002/jee.20393>
- Tan, P., Steinbach, M., & Kumar, V.** (2005). *Introduction to data mining*. Pearson.
- Vermunt, J. K., & Magidson, J.** (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 2, 549–553.
- Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., ... & Ziemba, E.** (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–30. <https://doi.org/10.15678/EBER.2023.110201>
- Waelen, R. A.** (2023). The struggle for recognition in the age of facial recognition technology. *AI and Ethics*, 3(1), 215–222. <https://doi.org/10.1007/s43681-022-00146-8>
- Weiss, H. M., & Rupp, D. E.** (2011). Experiencing work: An essay on a person-centric work psychology. *Industrial and Organizational Psychology*, 4(1), 83–97. <https://doi.org/10.1111/j.1754-9434.2010.01302.x>
- Woo, S. E., Jebb, A. T., Tay, L., & Parrigon, S.** (2018). Putting the “person” in the center: Review and synthesis of person-centered approaches and methods in organizational science. *Organizational Research Methods*, 21(4), 814–845. <https://doi.org/10.1177/1094428117752467>
- Xu, X., Wei, S., & Cao, Y.** (2023). Moving beyond the “international” label: A call for the inclusion of the (in) visible international engineering students. *Journal of Engineering Education*, 112(2), 253–257. <https://doi.org/10.1002/jee.20513>
- Yellamraju, T., Magana, A. J., & Boutin, M.** (2019). Investigating students’ habits of mind in a course on digital signal processing. *IEEE Transactions on Education*, 62(4), 312–324. <https://doi.org/10.1109/TE.2019.2924610>
- Zimmer, M.** (2010). “But the data is already public”: On the ethics of research in Facebook. In *The ethics of information technologies* (pp. 229–241). Routledge. <https://doi.org/10.1007/s10676-010-9227-5>
- Zuberi, T.** (2008). Deracializing social statistics: Problems in the quantification of race. In T. Zuberi & E. Bonilla-Silva (Eds.), *White logic, White methods: Racism and methodology* (pp. 127–134). Rowman & Littlefield.

TO CITE THIS ARTICLE:

Niu, J., & Reeping, D. (2024). Humanizing Data-Driven Methods in Engineering Education Research: A Systematic Literature Review of Four Journals From 2011 to 2021. *Studies in Engineering Education*, 5(2), 150–174. DOI: <https://doi.org/10.21061/see.159>

Submitted: 16 November 2023

Accepted: 12 September 2024

Published: 25 October 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Studies in Engineering Education is a peer-reviewed open access journal published by VT Publishing.