Cluster Analysis in Engineering Education

Mr. Andrew Jackson, Purdue Polytechnic Institute

Andrew Jackson is currently pursuing a PhD in Technology through Purdue's Polytechnic Institute, with an emphasis on Engineering and Technology Teacher Education. His research interests are engineering self-efficacy, motivation, and decision making. Andrew is the recipient of a 2015 Ross Fellowship from Purdue University and has been recognized as a 21st Century Fellow by the International Technology and Engineering Educators Association. He completed his Master of Science in Technology Leadership and Innovation at Purdue University with a thesis investigating middle school engineering self-efficacy beliefs. He previously taught middle school and undergraduate technology courses, accompanying both experiences with classroom research to improve practice.

Prof. Nathan Mentzer, Purdue University, West Lafayette (College of Engineering)

Nathan Mentzer is an assistant professor in the College of Technology with a joint appointment in the College of Education at Purdue University. Hired as a part of the strategic P12 STEM initiative, he prepares Engineering/Technology candidates for teacher licensure. Dr. Mentzer's educational efforts in pedagogical content knowledge are guided by a research theme centered in student learning of engineering design thinking on the secondary level. Nathan was a former middle and high school technology educator in Montana prior to pursuing a doctoral degree. He was a National Center for Engineering and Technology Education (NCETE) Fellow at Utah State University while pursuing a Ph.D. in Curriculum and Instruction. After graduation he completed a one year appointment with the Center as a postdoctoral researcher.

Cluster Analysis in Engineering Education

Abstract—This research paper describes cluster analysis methods and presents an example application of the clustering procedure for an introductory design class. Cluster analysis is a method developed from a diverse range of fields for empirically identifying groups among data. Despite its wide use, in engineering education research few examples of cluster analysis were identified in a content analysis of articles from the *Journal of Engineering Education*. We describe cluster analysis for the educational researcher in terms of 1) selecting features upon which to cluster, 2) computing similarities among cases, 3) clustering methods, and 4) calculating and verifying cluster results. To further introduce cluster analysis we present an example of using decision-making ratings to identify high-performing, moderate-performing, and low-performing design teams in our freshman design thinking course. We verify this cluster solution through ANOVA comparison to group project grade, conflict ratings, and satisfaction ratings. No significant differences were found for project grade, however group differences were observed for conflict ($p < .001$) and satisfaction ratings ($p < .001$).

Introduction

Arguably, the purpose of quantitative analysis is to explain events.[1] This process can occur through descriptive analysis – such as means, standard deviations, ranges – of what has been recorded from observation. It can be done through inferences, hypothesis testing, comparing groups to a model of some kind, or through goodness-of-fit indices which provide an array of criterion for decisions about the authenticity of analysis. In educational contexts, these procedures are used to "observe students' behavior and…to draw reasonable inferences about what students know" (p. 833)[2]. This information consequently informs instructors and other stakeholders in instructional processes.

A key element of these analyses are the underlying patterns or groupings in the data which researchers hope are informative. Using methods such as analysis of variance (ANOVA), *t* tests, or tests of invariance, these groupings are explicated by the researcher and are known a priori. However, *cluster analysis*, a method for identifying those groupings which are close together *solely from the data provided*, may prove useful in augmenting instructor and stakeholder understanding of student characteristics. "Cluster analysis is the organization of a collection of patterns into clusters based on similarity" (p. 265)[3] and can be useful for describing sets of entities, in our case students, based on their reactions on researcher specified variables. The successful application of cluster analysis may help technology and engineering educators fulfill the responsibility to advance personalized learning—one of the Engineering Grand Challenges (see Vest[4] and http://www.engineeringchallenges.org)—if it is used to identify what students know and do, and subsequently inform instructional interventions toward individualized student growth.

To provide a context and need for discussing cluster analysis generally, first, we characterize a limited number of articles published in the *Journal of Engineering Education (JEE)* which used the statistical approach. Next, we discuss steps in performing cluster analysis. Then, we will provide a brief description of researcher options while conducting a cluster analysis including several clustering algorithms which form a step-by-step procedure for performing clustering.

Finally, an example of clustering with data gathered from sections of an introductory design course will be described.

Content Analysis and Motivation

As an overarching term for a family of techniques, cluster analysis has many names [3, 5]; some relate to the emergence of clustering techniques from a variety of research fields nearly simultaneously, while others refer to the specific clustering approach being used. Aldenderfer and Blashfield[6] described a marked early growth in the number of publications on clustering methods in many disciplines just prior to the time of their writing. Clustering approaches have been used in biology, psychology, archeology, industrial engineering, marketing, computer vision, character recognition, machine learning and other fields, and are often use for exploratory pattern analysis and grouping when little is known about the data.

To situate a discussion of cluster analysis in the engineering education discipline, we performed a search for the terms "cluster" and "cluster analysis" within 24 volumes of the *JEE*, a leading engineering education research journal, available online. The search returned 139 articles. Upon reviewing the content of each article to characterize elements of cluster analysis reported on, only five articles reported using the statistical technique as opposed to using the term with some other meaning. (For example, we saw articles refer to clusters of core classes or clustering students together for collaborative learning; one article also reported the follow-up study to using cluster analysis and was not included.) Table 1 includes characteristics of these studies using a taxonomy for reporting cluster analysis informed by Clatworthy, Buick, Hankins, Weinman, and Horne.[7] Perhaps due to the relatively emergent status of engineering education research as a discipline,[9] few examples of clustering were identified. The scarce use of cluster analysis suggests the utility of our methodological introduction and example here.

Table 1. Characteristics of Five Cluster Analysis Studies Published in the *Journal of Engineering Education*

Characteristic Reported	Frequency
Software Used	2
Similarity Measure	2
Correlation	0
Euclidean	1
Mahalanobis	0
Clustering Method	5
Hierarchical/Agglomerative	3
Partitioning	4
Model Based	0
Density Based	0
Verification Method	4

Note. The total frequency of articles for each category may be more than the total as two articles repeated cluster analysis procedures to verify their solutions.

What service can cluster analysis provide in technology and engineering education? The goal of cluster analysis techniques is to reduce data to group of similar entities[10]; because clustering has

been applied in many research disciplines, the types of entities can be very diverse. Relative to educational research, the typical entity would be a student who is described by a number of variables contained in the data set. Among the *JEE* articles identified, cases included students[10-12] as well as engineering courses[13] and exam problems[14]. Identification of patterns is intuitive for humans, although we are limited in our capabilities. "Humans perform competitively with automatic clustering procedures in two dimensions, but most real problems involve clustering in higher dimensions. It is difficult for humans to obtain an intuitive interpretation of data embedded in a high-dimensional space" (p. 268)[3]. Because of the complexity of the information we obtain, an empirical method is necessary for use with large data sets, multiple variables (high-dimensions), and more objective consideration of the clusters. The researcher is able to focus on interpretation rather than identification of the clusters.[15]

The utility of grouping cases in our data is manifold. Aldenderfer and Blashfield[6] list several main goals:

1) development of a typology or classification,
2) investigation of useful conceptual schemes for grouping entities,
3) hypothesis generation through data exploration, and
4) hypothesis testing, or the attempt to determine if types defined through other procedures are in fact present in a data set. (p. 9)[6]

Cluster analysis is most commonly used for developing a classification because the results of a clustering solution identify the group assignment and can describe key characteristics of the groups. Sometimes, this development is exploratory; as will be discussed later, many of the decisions made in clustering tasks are ad hoc and made after considering multiple options. Cluster analysis may be used for hypothesis generation or testing, although this is not often conducted in a formal sense. In fact, it may simply be that the cluster solution aligns with results predicted from other research findings.[6]

Steps in Cluster Analysis

The typical cluster activity involves the following steps, consistent with Jain et al.[3,] Aldenderfer and Blashfield[6,] Jain and Dubes[15,] and Milligan and Cooper[16]: 1) selection of a sample to be clustered; 2) selection of a set of features upon which to cluster; 3) computation of similarities among cases; 4) completion of clustering or grouping; and 5) calculation of the resultant clusters, including assessment of the clustering. Some of the steps are completed noticeably while others are completed by statistical software during the process.[17] Each of these steps requires decision making and holds opportunities for divergent solutions.

Because sampling procedures are similar to other quantitative and qualitative methods they will not be discussed in detail. It is important however, that the individual entities sampled can produce meaningful clusters (i.e., we understand what a group of that entity means). Each of the next steps is discussed in turn, with several options and recommendations.

Selection of Features

Similar to other quantitative procedures, selection of variables for inclusion in the analysis plays a critical role in the resulting solution. For cluster analysis, further important decisions need to be

made regarding any transformation or standardization of variables. Researchers also need to reject the temptation to include all variables in the clustering procedure despite the heuristic definition given that clustering is useful for data reduction—this succumbs to "naïve empiricism" (p. 20)[6] and can produce incorrect clustering solutions.

Jain et al.[3] recommended "[isolating] only the most descriptive and discriminatory features in the input set, and [utilizing] those features exclusively in subsequent analysis" (p. 271)[3]. However implicit, some theoretical grounding should also support the clustering process and selection of variables; this theoretical rationale and the clustering solution can mutually fortify one another. These features may be based on selection of existing variables in the data set or extraction of new features through transformation of existing variables, as informed by theory. Feature extraction, or applying some transformation to sets of variables, can be used to reduce complexity of the data and especially to reduce the number of dimensions for needed visualizing the clustering solution.[3] For example, cluster plots are a common visualization that portray multiple variables reduced into principal components so they can be seen in two dimensions (for example, see [18] and Figure 6). "The feature selection process is of necessity ad-hoc, and might involve a trial-and-error process where various subsets of features are selected, the resulting patterns clustered, and the output evaluated" (p. 271)[3].

Computation of Similarities

This third step of a clustering task presents another researcher decision about the method for detecting similarities among the cases in the data set. Conversely, this might be termed identification of dissimilarities because in certain cases a greater result means that the data are at greater distances from one another (this is based on the measure utilized in the clustering procedure). Primary methods for similarities include correlation among the cases or distance measures, although other methods exist.[6] There are many distance measures, a few of which are described here.

1) *Correlation measures.* Correlation measures can be used as an indicator for similarity between two cases and the interpretation is similar to commonly used correlation (e.g., Pearson's r). Two correlated cases will move together on the profile of variables used in cluster analysis. A limitation of correlation coefficients as a similarity metric is that they do not take into account the distance of profile points; two cases with the same rises and falls on a set of variables could have significantly different levels on each variable but still be clustered based on their perfect correlation (see Figure 1).

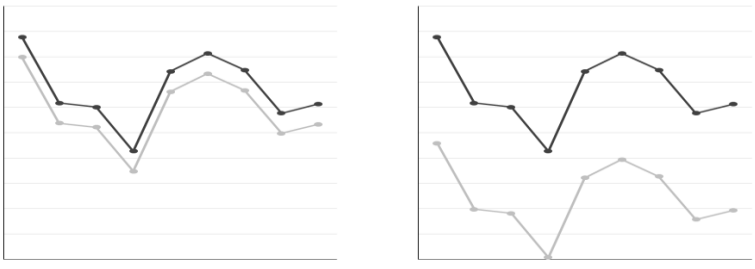

Figure 1. Example profiles of cases with perfect correlation. When using the correlation measure, distance is not considered and despite the visible dissimilarity between profiles on the right plot, entities would likely be assigned membership to the same cluster.

2) *Euclidean distance.* An intuitive measure for distance, Euclidean distance is the straight-line distance between two points in multidimensional space. It is calculated using

$$d_E(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

where *n* represents the number of dimensions or variables being considered and *p* and *q* represent the respective measures of two points on each of the dimensions. A variation on this distance is the Squared Euclidean distance which eliminates the square root operation on the formula to place greater weight on outlier cases. A result of Euclidean distance is that features with larger variance will tend to dominate the other variables; it is recommended to standardize the variables used to a mean of 0 and unit variance.[3, 6] Euclidean distances take into account all three dimensions of multivariate data—level, scatter, and shape.[5]

3) *Mahalanobis distance.* This distance is a weighted measure used in multivariate statistics that takes into account both the distance and covariance of the variables. If there is no correlation between the variables the Mahalanobis distance will equal the Euclidean distance.[15] Mahalanobis distance measures are reported to identify elliptical shaped clusters while Euclidean distances are effective for compact, isolated spherical clusters.[19]

The calculation of similarities, or dissimilarities depending on the measure, is often an intensive process because of the large number of calculations and comparisons which need to be made. This is multiplied by the number of cases in the data and number of variables selected for identifying clusters. More specifically, to work through a similarity matrix requires the evaluation of $n(n-1)/2$ unique values. For only 200 cases that is 19,900 values which represented a computationally difficult process prior to the ubiquity of computers used in computational analysis.[6]

Completion of Clustering

In cluster analysis using software packages, the computation of similarities or dissimilarities (Step 3) and the completion of clustering (Step 4) are often invisible phases of the process indicated only by the arguments and function calls to be executed. Completion of the clustering typically entails each case of data being assigned to a cluster with the final solution being stored or reported. (A few clustering algorithms operate differently here, with some algorithms excluding cases from a cluster if they do not fit well, and some algorithms providing probabilistic indicators for cluster membership.) A variety of clustering algorithms, will be discussed in the Clustering Methods section.

Calculation and Assessment of Clusters

Verification is an important step of cluster analysis because most clustering processes will produce clusters even when there are none.[3, 6, 17] Interpreting clustering solutions often involves comparing the stability of the cluster solution with other methods or number of clusters, using external criteria from those selected for comparing the clusters, or using internal measures of the cluster solution.[3, 16]

The stability of a cluster solution is an important consideration for its authenticity because if the cluster results are reproducible and robust to different approaches, we consider them unlikely to occur due to chance. Replicability with other samples reinforces the resulting cluster solution.[16] Clustering tendency measures can include comparison to randomized data in the same data space, or repetition of the analysis with cross-validation strategies. When using repeated analysis on the same data, stability is assessed by examining the agreement between clustering methods in the cluster assignment of cases.[13] If these new clustering tasks produced similar solutions to the original solution, the researchers have evidence of the cluster structure of their initial operation.

When verifying the solution with post hoc tests, variables used as features in the clustering will likely produce significant results, however these are *not meaningful* for validating a clustering solution because the distance has already been considered. These results aid interpretation of the cluster solution and suggest which variables have been most influential in determining the solution.[6, 17] Internal or external criteria can be used in the form of labels, which are useful for identifying a set of characteristics to automate the classification process in the future.[20] Comparison with external variables strengthens the interpretation of a cluster solution similar to the concept of criterion validity [21]; significance tests can then be performed using procedures appropriate for group comparison on the data with a cluster assignment serving as the group factor. Examples of this follow-up analysis could include ANOVA, the Kruskal-Wallis H test, or Pearson's chi-square test.

Internal measures of a clustering solution are based on follow-up calculations for the clusters. Tan et al.[20] lists cohesion within clusters, separation between clusters, and the silhouette coefficient, which considers the former two indices, as examples of internal criteria. For example, the silhouette coefficient is calculated by the following process [22]:

> For the ith object, let $a(i)$ be the average distance to all objects within its cluster, A.
>
> Then, let $d(i, C)$ be the average of all distances to all objects in another cluster, C.
>
> After computing all $d(i, C)$ for $C \neq A$, then $b(i) = min\{d(i, C)\}$.
>
> $$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$$

The silhouette coefficient for each point will range from -1 to 1 and indicates that a solution is weakly clustered when the average $s(i) \geq .26$, reasonably structured when the average $s(i) \geq .51$, and contains a strong structure when the average $s(i) \geq .71$.[23] Put simply, "the greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering" (p. 490)[20].

Clustering Methods

Classifying Cluster Approaches

"The variety of techniques for representing data, measuring proximity (similarity) between data elements, and grouping data elements has produced a rich and often confusing assortment of clustering methods" (p. 265).[3] In the interest of clarifying the array of clustering algorithms, introducing classifications is appropriate for understanding each approach, although different

cross-cutting classifications are available.[3] The specific clustering algorithm is used in Table 1 rather than these classifications which are implicitly tied to each algorithm. Clustering approaches can be described as Hierarchical (Agglomerative) or Partitioning (Divisive); Hard (Exclusive) or Fuzzy (Fractional or Probabilistic Membership, Non-Exclusive) or Overlapping (Non-Exclusive); or Deterministic or Stochastic. General descriptors for each of these methods are given, followed by a more specific discussion of some clustering approaches.

1) *Hierarchical versus Partitioning.* Hierarchical algorithms produce a cluster solution of nested clusters ranging from all cases being their own cluster to all cases being subsumed as one cluster. The trajectory of cluster merging is reported for the researcher to inspect and identify a cut point. On the other hand, partitioning algorithms begin at an intermediate stage by assigning cases to clusters. This assignment is typically based on starting points for a predicted number of clusters supplied by the researcher. "Hierarchical algorithms…are useful for exploratory work when researchers do not have a preconceived idea about the likely number of clusters in the dataset. Non-hierarchical algorithms, on the other hand, are appropriate when there is a theoretical or empirical rationale for predicting the number of clusters or when the data set is large" (p. 285),[17] although these are just rules of thumb.

   One argument for this may be that visualization of hierarchical methods will display all of the nested factor solutions and allow the researcher to define the number of clusters. Partitioning approaches typically require that the researcher specify the number of cluster patterns to search for in the data; this approach seems to align with the process of confirming an expected cluster pattern.[17] This heuristic for algorithm selection provides guidance but doesn't exclude the researcher from attempting runs of the partitioning algorithm with varying parameters (i.e., trying two, three, or four clusters) and comparing the resulting solutions.

2) *Hard versus Fuzzy versus Overlapping.* In hard clustering algorithms, each entity of the data is assigned to be in a cluster in the final solution. In fuzzy algorithms, a matrix is produced that reports the probability of membership in each cluster for each case. In overlapping cluster solutions, people may be assigned to more than one cluster when their membership is unclear. Hard clustering can be viewed as a special case of fuzzy clustering where cluster assignments are made on the basis of the strongest probability for each case. In practice, hard clustering solutions are most common.

3) *Deterministic versus Stochastic.* Deterministic solutions are those where no randomness is involved. Deterministic techniques "guarantee an optimal partition when performing exhaustive enumeration" (p. 288).[3] Hierarchical techniques are all deterministic. Stochastic techniques are affected by some element of randomness. In the case of partitioning algorithms, for example, random starting points are often initiated for the clustering algorithm to work from. These starting points can affect the cluster solution but "stochastic search techniques generate a near-optimal partition reasonably quickly" (p. 288).[3] In the case of a large amount of data, stochastic algorithms will operate with fewer resources. However, resulting cluster solutions may be slightly different unless the underlying pattern of cases is clear.

Clustering Algorithms

1) *Agglomerative.* This collection of clustering methods is directly related to the hierarchical classification just described. While each hierarchical method produces *n* possible solutions (from one encompassing cluster to each case as its own cluster), they differ in their criteria for merging cases and clusters. Several common agglomerative techniques include single linkage, complete linkage, centroid, and minimum variance algorithms.[20] In order, these algorithms will join cases based on the closest single connection, closest furthest point within clusters, closest center point of a cluster, or minimal sum of squared distance value that can be produced. These solutions tend to find spherical clusters, although single linkage clustering can identify unique shapes or elongated patterns if the data are close together. This can be a benefit or downside depending on the form of the data; if non-spherical shapes exist, single link clustering may be appropriate to identify these patterns.[3] A cursory visual of these methods is provided in Figure 2.

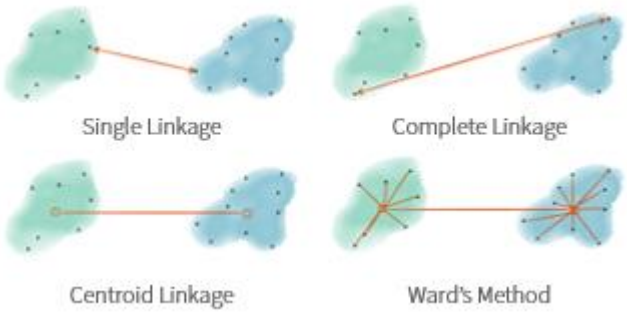

Figure 2. Visual depiction of agglomerative clustering methods. Most tend to find spherical shapes, although single linkage can overcome this limitation.

2) *Partitioning.* K-means clustering is the most common method for partition cluster approaches. A researcher specified number of clusters, *k*, is entered into the algorithm which then iteratively assigns points to a nearest starting centroid, and then recalculates a more accurate centroid for the cluster. The process repeats until no reassignment of cases is made.[3, 24] K-means analysis will seek compact, spherical clusters of similar densities.[20]

3) *Model Based.* Model based clustering approaches, also called mixture resolving approaches[3], are less commonly used, however present an opportunity to address potential limitations of other clustering methods. These approaches are based on the assumption that underlying patterns of data are drawn from a distribution, most often Gaussian. The parameters of the distribution are estimated and the algorithm iteratively refines the parameter estimates with maximum likelihood processes—identifying the most likely parameters based on the data observed.[3] An important specification is the type of distribution to be used in estimating the clusters, which will affect the cluster solution. Model based clustering is accommodating to clusters of different densities however it requires model specifications be determined by the researcher. Figure 3 depicts a model based cluster example in comparison to density based clustering.

4) *Density Based.* A final form of clustering algorithm which is described here is based on the proximity of points within a multidimensional space. DBSCAN, a popular algorithm for density based clustering attempts to parse out core, border, and noise points from the data.[20] Two parameters are determined by the researcher: the number of nearest neighbors, and the maximum distance.[26] If a point exceeds the minimum number of neighbors within the threshold distance, it is identified as a core point. Each successive point in the data set is evaluated and a cluster solution identified. Border points are those that are within the specified distance of core point yet do not have the specified number of nearest neighbors; noise points are not within the specified distance of a core point. Density based methods will only separate distinct clusters surrounded by low-density space. Like model based approaches, the search specifications are up to the user and therefore can create difficulty in identifying appropriate solutions. Misspecification of either parameter can cause the algorithm to identify only noise or to generate one cluster.

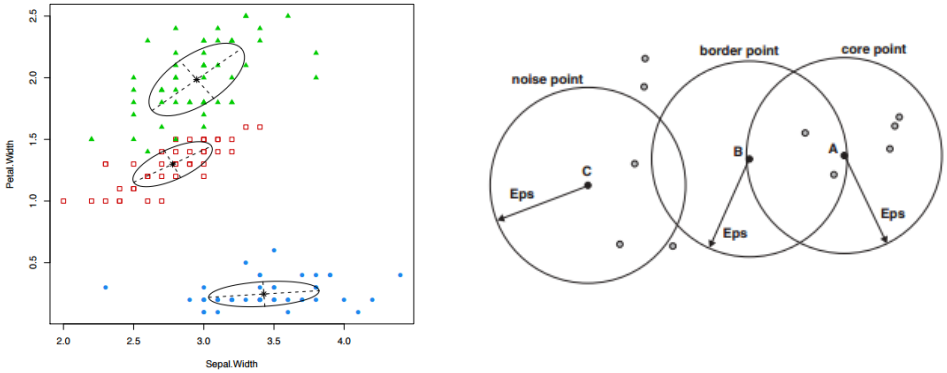

Figure 3. Model based and density based clustering representation. Images from Fraley et al.[25] and Tan et al.[20], respectively, show the model-based distribution being identified, and core, border, and noise points being identified.

Summary

Each of the clustering algorithms presented takes a different approach in solving the problem of identifying patterns or groups among data. Because of the complications associated with this empirical process (e.g., it is resource intensive, at times subjective, and each approach has limitations), the educational researcher needs to take care when applying this strategy to be sure that the results meaningfully contribute to understanding of student attitudes and achievement.

Next we present an application of partitioning clustering methods including a theoretical brief, research question, and description for research methods. In the results section, several visualizations for clarifying the clustering solution are used. This application is presented as an example, despite concerns about the form of data which will be described.

A K-Means Clustering Example

Within engineering design activities, the development of teamwork skills is frequently cited as a goal for instruction.[27, 28] Specifically, the ability to make decisions on effective designs is critical among these teamwork skills. This decision-making ability was identified by Mentzer[29] as a limitation among beginning designers. As part of an instructional team for a freshman design course, we attempted to measure the quality of student group decision processes to improve future

instruction. This proved difficult because an evaluation instrument for group decision processes in the context of engineering design projects was not found.[30] A 15 question self-reflection instrument was developed to facilitate measurement of the group design process. Questions pertained to three factors of decision making: processing information, collectively understanding decisions, and considering alternatives.

Theoretically, these underlying factors of group decision-making will have an effect on team dynamics.[31] Prior research has identified a negative relationship between team conflict and group decision making.[31, 32] This relationship is explained by the negative outcomes of conflict toward information sharing and team cohesion. On the other hand, satisfactory outcomes of effective decisions might include positive affective beliefs toward team members. Another good indicator for the accuracy of the instrument would be an association with course or assignment grades related to the decision-making process under consideration. While student perceptions of the group decision-making process are important on their own, these outward manifestations of decision-making consequences would support future instructional work in this area. Using evidence from the decision making instrument, these conceptual relationships are untested, as of yet. We ask, "Can cluster analysis help us identify clusters of students with similar perceptions on their decision-making process?" Also, "Will these resulting student groups be related to external criteria theoretically aligned with decision-making processes?" These comparisons may work to substantiate a cluster solution and the group decision making instrument. We used k-means clustering as the data was exploratory and we were unsure what the underlying cluster structure of students by decision-making score would be.

For this work, data from two semesters were collected as part of an end-of-semester reflection following participation in a team based design activity. Students identified a variety of decisions which took place in their projects, and were asked to use the decision making instrument to evaluate the group decision-making process of their team. Other artifacts from the course were collected for external validation of cluster solutions. These included: team satisfaction ratings obtained from the CATME system for team evaluation[33]; conflict perceptions based on questions adapted from Amason[32] and Dean and Sharfman[31]; and student percentage score on the final project. All procedures were conducted with R software (Version 3.2.2)[34].

Results

Prior to in-depth analysis related to underlying cluster patterns in the data, features were carefully considered for selection in the k-means analysis. Feature selection was guided by the theories of group decision-making and its antecedents, as well as the psychometric properties of the variables used. For student responses from the semesters used, the decision-making instrument had strong correlations among all of the questions. This is not surprising since the scale structure has been shown to contribute to a second-order factor for decision making effectiveness.[30] Each of the subscales maintained strong internal consistency, as indicated by Cronbach's α ranging from .84 to .91. Next, dependent variables were evaluated similarly and findings were supportive of the consistency of these scales.

For cluster analysis, cases with missing decision-making variables were removed (because similarity measures cannot be calculated on the basis of missing variables). This left 753 observations (80% of the original set). The decision making questions were scaled to have a mean

of 0 and unit variance; this ensures that one of the indicators does not overcome the influence of another if it has greater variance.[6]

Determining the number of clusters. In order to determine the number of clusters, two visual tools were used which are based on criterion evidence of various cluster solutions. First, the within cluster sum of squares (WSS) was calculated for cluster solutions containing one to 15 clusters. The plot of these WSS values revealed a bend at 3 or 4 clusters, suggesting that the addition of more clusters was not useful in reducing the error of the data (see Figure 4).

Next, the average silhouette width[22] was used to illustrate the internal quality of various cluster structures ranging from 1 to 10. This was done with the "fpc"[35] package in R. The criterion suggested different results, using only a two cluster solution (see Figure 5). Because of the contrasting results, a two and three cluster solution were explored visually once more, using cluster plots (see Figure 6). Cluster plots extract principal components in an attempt to represent high-dimensional space in two dimensions. Because 15 variables were used in this analysis, the reduction still showed overlap in the data, however a three cluster solution explained more variance within clusters (48.5% compared to 39.5%) and was chosen for future description and verification.

Using the k-means method which implicitly uses Euclidean measures to reduce WSS (since the distance between the centroid is iteratively reduce), the three cluster solution was fairly balanced with 318, 297, and 138 students in each group. Attributes of each cluster were identified by comparing scores on each decision-making factor (processing information, collectively understanding decisions, and considering alternatives) to the overall means.[12] The cluster centers provide a clear profile for each cluster with the first cluster having above average performance for each decision making item, the second cluster having slightly below average performance, and the third cluster having poor performance on each item (see Table 2).

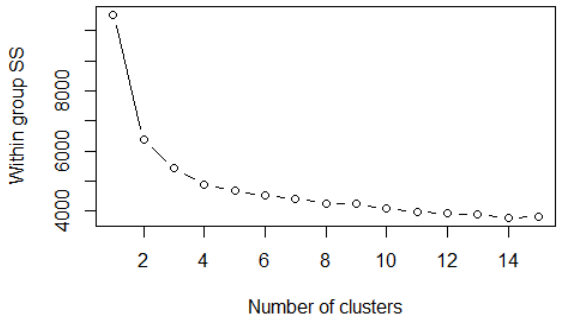

Figure 4. Plot showing the reduction of within cluster sum of squares (WSS). A bend in the plot suggests a useful cluster solution.

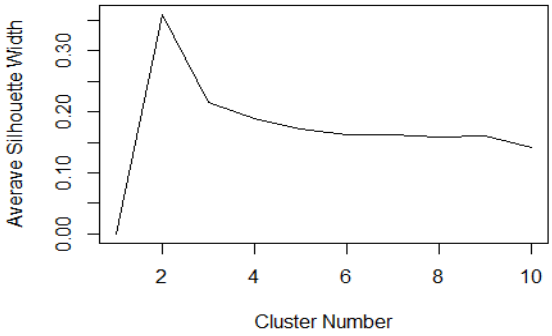

Figure 5. Average silhouette width for one to 10 clusters. The peak in the plot suggests that the cohesion and separation of clusters is maximilzed with a two cluster solution.

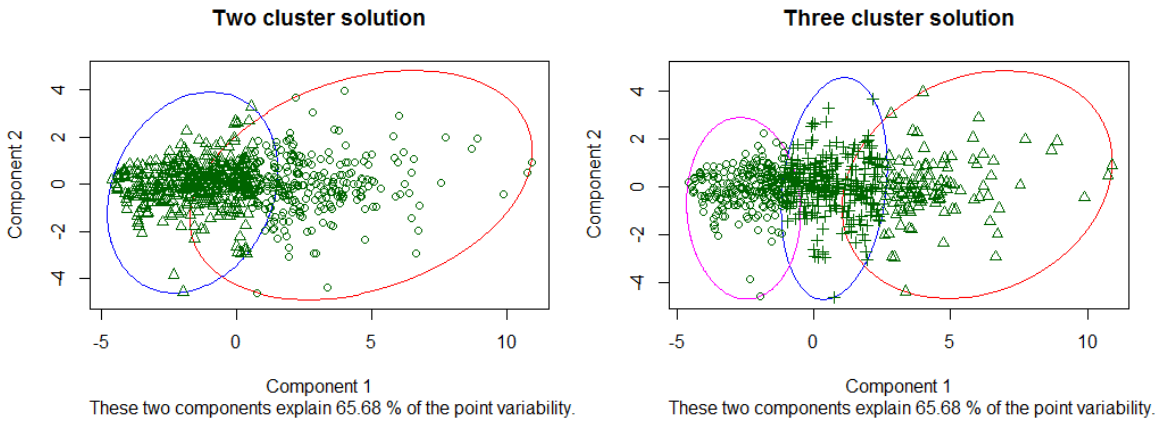

Figure 6. Comparison of the two and three cluster solutions on a cluster plot.

Table 2. Cluster Centers and Comparisons

	Cluster 1 High performing	Cluster 2 Moderate performing	Cluster 3 Low Performing
Decision Factors		Cluster Centers	
Information (*M* = 5.50)	6.34	5.26	4.07
Alternatives (*M* = 5.47)	6.33	5.21	4.04
Understanding (*M* = 5.51)	6.26	5.34	4.15
		Cluster Comparisons	
Project %	91.23%	91.05%	90.86%
Conflict	2.65	2.73	**3.56***
Satisfaction	**4.63***	4.4	4.24

* p < .001 difference from other groups

Verification with external criteria. The external variables included in the data set were compared to the clustering solution using one-way ANOVA tests for each variable. When the results were significant, Tukey Tests were administered for pairwise comparison. There was not a significant relationship between cluster group and final project grade percentage, $F(2, 750) = .184$, $p = .832$. This suggests that despite struggling within groups on their decision-making process, student assignments were completed and graded without effect. This is perhaps due to the compilation of group effort that went into each assignment, or the aggregation of student final project milestones into a single criterion for this analysis. In the future, narrowing this comparison to decision-making assignments of the final project may show an effect on group performance.

Within team conflict did have a significant difference by cluster, with the teams performing high and moderate on decision-making having significantly less conflict than the poor decision-making team, $F(2, 750) = 15.36$, $p < .001$. Team satisfaction was also significantly different by team, with the team reporting high decision-making having significantly higher satisfaction than the other two, $F(2, 750) = 14.54$, $p < .001$. Aggregated outcome variables for these external criteria are included in Table 2.

Limitations and Recommendations

While limitations reside with each of the clustering methods previously described, there are concerns for the approach holistically as well. There is not a well agreed upon set of rules for determining the number of clusters.[16] Selecting an appropriate number of clusters to draw from the data is difficult because of the subjective nature, various indices which might be used, and conflicting results which can occur if the model is specific with different arguments (e.g., the number of clusters, similarity measure, etc.). Cluster analysis results are sensitive to the distance measures and algorithms used.[17] This requires that the selection of these parameters be done carefully and verification procedures be undertaken to ensure that the cluster solution is meaningful. "Results of cluster analysis should be triangulated and tested in a variety of ways" (p. 396)[17] to ensure that the clustering solution is valid. As described in *Calculation and Assessment of Clusters* section previously, researchers have several methods for conducting this evaluation.

Several concerns regarding emergent clusters are identified in addition to the aforementioned concern that clustering algorithms will produce a solution regardless of the actual form of the data. Egan[36] argued that identification of a clustering structure using colorful descriptions and stereotypes can discourage researchers from looking more closely at the data and that widespread use could result in "irreducible typologies" (p. 152)[36]. It is important to remember that cluster analysis is descriptive rather than explanatory and "at best it is a convenient summary of findings on individual variables" (p. 152)[36]. While clustering presents a useful summary of the groups among our data, it is only useful insofar as it is an authentic representation of the actual patterns of our data. Like with other statistical methods, the planned application and implications of our conclusions need to align with the validity examinations conducted in our work.[37]

Specifically, for the description of these decision-making profiles among beginning design students, the confusion between two or three clusters in the solution, and visual inspection of the data suggests that further verification is would be helpful. Nevertheless, our work is informative for demonstrating steps cluster analysis as a technique. We have reported the software used (R); distance measure (Euclidean distance) and algorithm used (k-means); and verified our cluster solution through comparing multiple solutions and using internal and external criteria. Further work needs to be done to evaluate student decision-making upon the external criteria described here, and perhaps identify different student attitudes based on these results.

There are certainly diverse students enrolled in our classes, with unique characteristics. Employing cluster analysis does not seek to ignore these individual identities, but rather, complement our view as instructors by identifying student groups that are similar upon features that we determine are important. As noted by Nelson et al.[11], this approach "may help engineering educators better tailor their courses, and especially their foundational courses" (p. 95). The follow-up to these analysis can help our students feel greater personalization in their instruction as we tailor our teaching to their needs.

References

1. Aron, A., Aron, E., & Coups, E. J. (2009). *Statistics for psychology* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
2. Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831-841. doi: 10.1002/tea.21032
3. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323. doi: 10.1145/331499.331504
4. Vest, C. M. (2008). Context and challenge for twenty-first century engineering education. *Journal of Engineering Education, 97*(3), 235-236. doi: 10.1002/j.2168-9830.2008.tb00973.x
5. Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456-473. doi: 10.1037/h0057173
6. Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Thousand Oaks, CA: SAGE Publications.
7. Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British journal of health psychology, 10*(3), 329-358.
8. Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838.
9. National Engineering Education Research Colloquies Steering Committee. (2006). The research agenda for the new discipline of engineering education. *Journal of Engineering Education, 95*(4), 259-261. doi: 10.1002/j.2168-9830.2006.tb00900.x
10. Carpenter, S. L., Delugach, H. S., Etzkorn, L. H., Farrington, P. A., Fortune, J. L., Utley, D. R., & Virani, S. S. (2007). A knowledge modeling approach to evaluating student essays in engineering courses. *Journal of Engineering Education, 96*(3), 227-239. doi: 10.1002/j.2168-9830.2007.tb00932.x
11. 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. *Journal of Engineering Education, 104*(1), 74-100. doi: 10.1002/jee.20066
12. Lin, C.-C., & Tsai, C.-C. (2009). The relationships between students' conceptions of learning engineering and their preferences for classroom and laboratory learning environments. *Journal of Engineering Education, 98*(2), 193-204. doi: 10.1002/j.2168-9830.2009.tb01017.x
13. PÉRez, C. D., Elizondo, A. J., GarcÍA-Izquierdo, F. J., & Larrea, J. J. O. (2012). Supervision typology in computer science engineering capstone projects. *Journal of Engineering Education, 101*(4), 679-697. doi: 10.1002/j.2168-9830.2012.tb01124.x
14. Kumsaikaew, P., Jackman, J., & Dark, V. J. (2006). Task relevant information in engineering problem solving. *Journal of Engineering Education, 95*(3), 227-239. doi: 10.1002/j.2168-9830.2006.tb00895.x
15. Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
16. Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied psychological measurement, 11*(4), 329-354.
17. Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*(3), 383-398. doi: 10.1007/s11423-012-9235-8
18. Kirn, A. N. (2014). *The influences of engineering student motivation on short-term tasks and long-term goals.* (Doctoral Dissertation), Clemson University, Clemson, SC.

19. Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (hec). *Neural Networks, IEEE Transactions on, 7*(1), 16-29. doi: 10.1109/72.478389
20. Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (1st ed.). Boston: Pearson Addison Wesley.
21. Barab, S. A., Bowdish, B. E., & Lawless, K. A. (1997). Hypermedia navigation: Profiles of hypermedia users. *Educational Technology Research and Development, 45*(3), 23-41.
22. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi: http://dx.doi.org/10.1016/0377-0427(87)90125-7
23. Spector, P. (2011). Cluster analysis. *Concepts in Computing with Data*. from http://www.stat.berkeley.edu/~s133/Cluster2a.html
24. Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*(1), 100-108. doi: 10.2307/2346830
25. Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation*.
26. Hahsler, M. (2015). Dbscan: Density based clustering of applications with noise (dbscan) and related algorithms (Version 0.9-5) [R package]. Retrieved from https://cran.r-project.org/package=dbscan
27. National Academy of Engineering. (2005). *Educating the engineer of 2020: Adapting engineering education to the new century*. Washington, DC: National Academies Press.
28. Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.
29. Mentzer, N. (2014). Team based engineering design thinking. *Journal of Technology Education, 25*(2), 52-72.
30. Jackson, A., Mentzer, N., & Zissimopoulos, A. N. (2015, June). *Factors of group design decision making.* Paper presented at the 2015 ASEE Annual Conference & Exposition, Seattle, WA. doi:10.18260/p.24098
31. Dean, J. W., Jr., & Sharfman, M. P. (1996). Does decision process matter? A study of strategic decision-making effectiveness. *Academy of Management Journal, 39*(2), 368.
32. Amason, A. C. (1996). Distinguishing the effects of functional and dysfunctional conflict on strategic decision making: Resolving a paradox for top management teams. *Academy of Management Journal, 39*(1), 123-148. doi: 10.2307/256633
33. Ohland, M. W., Bullard, L. G., Felder, R. M., Finelli, C. J., Layton, R. A., Loughry, M. L., . . . Woehr, D. J. (2005). CATME. West Lafayette, IN.
34. R Core Team. (2015). R: A language and environment for statistical computing. (Version 3.2.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
35. Hennig, C. (2015). Fpc: Flexible procedures for clustering (Version 2.1-10) [R package]. Retrieved from https://cran.r-project.org/package=fpc
36. Egan, O. (1984). Cluster analysis in educational research. *British Educational Research Journal, 10*(2), 145-153. Retrieved from: http://www.jstor.org/stable/1500750
37. Douglas, K. A., & Purzer, Ş. (2015). Validity: Meaning and relevancy in assessment for engineering education research. *Journal of Engineering Education, 104*(2), 108-118. doi: 10.1002/jee.20070