Utilizing Cluster Analysis of Close-Ended Survey Responses to Select Participants for Qualitative Data Collection

Ms. Katherine M. Ehlert, Clemson University

Katherine M. Ehlert is a doctoral student in the Engineering and Science Education department in the College of Engineering, Computing, and Applied Sciences at Clemson University. She earned her BS in Mechanical Engineering from Case Western Reserve University and her MS in Mechanical Engineering focusing on Biomechanics from Cornell University. Prior to her enrollment at Clemson, Katherine worked as a Biomedical Engineering consultant in Philadelphia, PA. Her research interests include identity development through co and extra-curricular experiences for engineering students.

Dr. Courtney June Faber, University of Tennessee

Courtney is a Research Assistant Professor and Lecturer in the College of Engineering Honors Program at the University of Tennessee. She completed her Ph.D. in Engineering & Science Education at Clemson University. Prior to her Ph.D. work, she received her B.S. in Bioengineering at Clemson University and her M.S. in Biomedical Engineering at Cornell University. Courtney's research interests include epistemic cognition in the context of problem solving, researcher identity, and mixed methods.

Dr. Marian S. Kennedy, Clemson University

Marian Kennedy is an Associate Professor within the Department of Materials Science & Engineering at Clemson University. Her research group focused on the mechanical and tribological characterization of thin films. She also contributes to the engineering education community through research related to undergraduate research programs and navigational capital needed for graduate school.

Dr. Lisa Benson, Clemson University

Lisa Benson is a Professor of Engineering and Science Education at Clemson University, with a joint appointment in Bioengineering. Her research focuses on the interactions between student motivation and their learning experiences. Her projects involve the study of student perceptions, beliefs and attitudes towards becoming engineers and scientists, and their problem solving processes. Other projects in the Benson group include effects of student-centered active learning, self-regulated learning, and incorporating engineering into secondary science and mathematics classrooms. Her education includes a B.S. in Bioengineering from the University of Vermont, and M.S. and Ph.D. in Bioengineering from Clemson University.

Utilizing Cluster Analysis of Close-Ended Survey Responses to Select Participants for Qualitative Data Collection in Mixed Methods Studies

Introduction

The purpose of this research paper is to discuss the application of cluster analysis within an engineering education mixed methods study to compare three clustering techniques (k-means, Ward's, and Complete Link), and then discuss the impact of this analysis on the process of selecting participants for interviews.

Cluster analysis involves the use of statistical methods to find groups within a set of data based on provided measures (Kaufman & Rousseeuw, 1990). This technique has been used within other disciplines to group different types of data and entities of systems, such as chemicals (Maccuish & Maccuish, 2014), manufacturing decisions (Lorentz, Hilmola, Malmsten, & Srai, 2016), or planets (Jiang, Yeh, Hung, & Yang, 2006), based on a series of factors or variables. In engineering education, cluster analysis has been used to group participants who have similar attributes such as epistemic beliefs (Faber, Vargas, & Benson, n.d.), activities within a learning environment (Antonenko, Toy, & Niederhauser, 2012; Galloway & Bretz, 2015a, 2015b), relative risk of attrition (Chan & Bauer, 2014), or who exhibit certain behaviors in courses (Karabenick, 2003; Raker et al., 2015; Shell & Soh, 2013; Stewart, Miller, Audo, & Stewart, 2012). Cluster analysis can help researchers who are using a mixed methods approach select participants for interviews when a variation of participant attributes or perspectives is desired. However, it is important to understand the underlying assumptions, boundary conditions, and limitations for cluster analysis, because the choice of algorithm will impact the type of clusters formed and the individuals assigned to those clusters. Within the context of this paper, we use cluster analysis to group participants based on closed-ended survey responses measuring different facets of engineering epistemic beliefs.

Cluster Analysis Overview

To ensure clarity throughout the paper, we first will define many of the terms that we will be using: population, participant, data set, data point, items, and factors. We refer to population in the statistical sense of all the individuals that meet our standard of criteria. For us, this is every student that received a recruitment email to participate in our survey. Anyone who choose to participate in the survey is called a participant. The data set is the collection of information that we have from our participants. A data point represents the participant during our data analysis and in the plots that we've provided. Items are the individual questions that the participants were asked on the survey. We grouped similar items together to create factors and calculated an average score for each factor.

Cluster analysis is often used in exploratory work where researchers are uncertain of the number of groups (clusters) within a data set. This grouping technique is best used to divide the data into smaller groups (clusters) that have similar characteristics across a select number of dimensions. Lattin, Carroll, and Green most accurately describe cluster analysis as "undertaken with the objective of addressing the heterogeneity of the data… explicitly divid[ing] the [data set] into more homogenous subsets" (p.264, 2003).

While cluster analysis is not a synonym for factor analysis, it does share some similarities. Antonenko, Toy, and Niderhauser relate cluster analysis "as complementary to factor analysis: factor analysis groups *variables* across cases (e.g. individuals), clustering algorithms group *cases* based on the variables of interest" (p. 384, 2012, emphasis in original). In other words, in factor analysis, the interest is in how variables group together and measure the same factor; in cluster analysis, the interest is in how individuals (cases) group together based on behaviors, beliefs, or other characteristics of interest.

One major caveat to consider when applying cluster analysis is that the information gathered from cluster analysis alone is not sufficient to determine if there are actual clusters in the population that is represented by the data set. Cluster analysis, at its core, is a set of optimization algorithms that will provide the optimal solution for the combination of the input data set and the clustering algorithm. As such, each algorithm will provide a solution, regardless of whether actual clusters exist. To determine if actual clusters exist, additional measures of the spread of the data (i.e. local density analysis) should be used.

The general process of cluster analysis begins with determining where the data are located relative to one another. This is done by using a proximity measure which calculates the distance between each piece of data relative to the variables of interest. The most common proximity measure is squared Euclidian distance which provides higher weight to larger distances and can be calculated beyond three dimensions. Another commonly used metric of proximity is the within-cluster sum of squares or the variability within a cluster. After the proximity measure is determined, the clustering algorithm is used to group similar data points together. Clustering algorithms mostly differ on their methods for grouping data, their robustness to outliers, and computational efficiency.

Two types of clustering algorithms are commonly used in data analysis: hierarchical and partitioning algorithms. Hierarchical algorithms typically begin with placing each data point in a separate 'cluster' and then pairing nearby clusters (with low proximity values) together until a single cluster exists. Criteria evaluating multiple clustering solutions (described in detail below) are analyzed and an optimal number of clusters is determined. Partitioning algorithms separate data points into a pre-specified number of clusters (partitions) and place data points into each cluster such that data points within the cluster are similar and data points outside the cluster are dissimilar. Again, criteria for multiple solutions are compared and an optimal number of clusters is determined. In this paper, we discuss selecting three algorithms are to be used in clustering our participants. The data set consisted of average scores of the participants' response to closed-ended survey items probing their engineering epistemic beliefs. The three algorithms we selected include two hierarchical (Ward's and Complete Link) and one partitioning (k-means). Each algorithm will use the same data set to show how the use of a method can affect the results.

We selected the three algorithms based on anticipated cluster behavior. Prior to clustering, we plotted the data in the first two principal components to visualize its behavior. The two primary principal components are the axes that show the most amount of variability in the data. In some areas of the plot, data points were clumped together and in other areas the data points were spread apart. To mitigate the effects of the data behavior on the cluster shape, we selected clustering techniques that would be less sensitive to the clumping: Ward's, Complete Link, and k-means.

Ward's Clustering Algorithms

Ward's algorithm systematically combines the clusters that, when merged, contribute the smallest increase to the within-cluster sum of squares (i.e. the group variability). Because cluster variability is affected by the number of data points within the cluster and their relative distance to the cluster centroid, Ward's method tends to produce clusters that have a similar number of data points and are relatively spherical. Spherical cluster shapes indicate that the variables being analyzed are equally important. Oblong cluster shapes would indicate that one or two variable(s) dominates the data behavior.

Complete Link Clustering Algorithm

The Complete Link algorithm, also known as the "farthest neighbor" (Rencher, 2002) algorithm, determines the furthest Euclidean distance between clusters and combines the two clusters that have smallest distance from the data points that are furthest apart. By using the data points that are furthest from each other in the clusters, Complete Link "ensures that each [data point] added to a cluster is close to all [data points] in the cluster and not just one" (p. 282, Lattin et al., 2003). Because the Complete Link algorithm evaluates data points that are the furthest apart, it is more sensitive to outliers than other methods.

K-means Algorithm

K-means is the most commonly used partitional algorithm. Although there are other partitioning algorithms, they are typically used for very specific partitioning needs and are not robust to a wide range of data sets. The k-means algorithm divides the space into *k* number of groups and concurrently minimizes the within-group variability and maximizes the between-group variability. The algorithm typically begins by randomly sectioning the space to the pre-defined groups and calculates the centroid of each group. Then evaluates the distance of each data point relative to the centroids. The data point is then assigned the group that it is closest to (smallest distance from the centroid). After all the data points are assigned to groups, the centroids are recalculated and the process is repeated. The repetition continues until there are no changes to the groups. Because the number of groups (i.e. partitions) is pre-determined by the user, k-means is best used when there is theoretical support for the number of groups selected. Since the initial centroids are randomly selected, the algorithm should be run multiple times to ensure robustness in the solution.

Summary of Mixed Methods Study

This work is situated within a larger, explanatory mixed-methods project focused on understanding how undergraduates conceptualize their identities as researchers and their engineering epistemic beliefs. The first phase of the project was a survey with closed-ended and open-ended items to probe students' perceptions of themselves as researchers, their contributions to their field, their beliefs about knowledge and where it comes from, and their need for cognitive closure. The second phase entails in-depth interviews of selected participants from the survey respondents to understand their beliefs and views on research, their researcher identity,

and epistemic beliefs. We will select interview participants based partly on the results of the cluster analysis of survey data (responses to close-ended questions).

Survey Design and Participant Population

While open and closed-ended items were on the survey, for the cluster analysis we only used the responses to the closed-ended items. We selected closed-ended items from previous studies to represent the following six factors: (1) Closed-mindedness (Webster & Kruglanski, 1994), (2) Discomfort with Ambiguity (Webster & Kruglanski, 1994), (3) Certainty of Knowledge (Yu & Strobel, 2012), (4) Sources of Knowledge (Yu & Strobel, 2012), (5) Simplicity of Knowledge (Greene, Torney-Purta, & Azevedo, 2010; Yu & Strobel, 2012), and (6) Justification of Knowledge (Ferguson & Braten, 2013; Greene et al., 2010; Yu & Strobel, 2012). Two factors (Closed-mindedness and Discomfort with Ambiguity) were used to measure a participant's need for cognitive closure (Webster & Kruglanski, 1994). The other four factors were used to probe a participant's epistemic beliefs in engineering (Ferguson & Braten, 2013; Greene et al., 2010; Yu & Strobel, 2012). Table 1 shows the six factors with an example item in the factor. All forty-five items and the factors with which they are associated are listed in Appendix A. Prior to the work discussed in this paper, we conducted a pilot study that probed participant's interpretations of item wording to help formalize item wording and ensure the participants were interpreting the items similarly (Faber et al., n.d.).

Table 1: Summary of factors that were evaluated during cluster analysis with an example item. The factors with associated items are all described in Appendix A.

Factor	**Example Item Within Factor**	**Reference that Originally Used Item**
Closed-Mindedness	I dislike questions which could be answered in many different ways.	Webster & Kruglanski, 1994
Discomfort with Ambiguity	I feel uncomfortable when someone's meaning or intention is unclear to me.	Webster & Kruglanski, 1994
Certainty of Engineering Knowledge	Theories in engineering cannot be argued or changed.	Yu & Strobel, 2012
Sources of Engineering Knowledge	You can trust the information you find in engineering textbooks.	Yu & Strobel, 2012
Simplicity of Engineering Knowledge	Engineers can solve engineering problems by just following a step-by-step procedure.	Yu & Strobel, 2012
Justification of Engineering Knowledge	I believe everything I learn in my engineering classes.	Greene, Torney-Purta, & Azevedo, 2010

Participants responded to the forty-five anchored questions on a 7-point scale (1 – strongly disagree, 7 – strongly agree). Prior to data analysis, negatively worded items were reverse coded to ensure alignment within each factor. We performed all data analysis using the statistical analysis software R (version 3.3.1) (Team, 2016) using the psych, fpc, cluster, car, psy, rgl, and MASS packages.

The survey was distributed to Mechanical Engineering (ME) and Biomedical Engineering (BME) students at five institutions varying in size, type, and location within the United States. Department representatives, such as undergraduate coordinators, in ME and BME departments distributed the survey through email to potential participants in their departments. The survey was reviewed and approved by University IRBs prior to its distribution in the Spring 2016 semester. Of the 113 students who began the quantitative portion of the survey, 109 students answered every question. 59.3% of the students were male, 40.7% female. Racial distribution was 68.1% Caucasian, 1.8% Black/African American, 12.4% South Asian, 11.5% East Asian, 2.7% Other Asian, 1.8% American Indian/Alaskan Native, and 11.5% identified as 'other'. Most respondents were from the two institutions with the largest total number of enrolled engineers (65.4%), with most of the responses coming from the largest institution (50.4%). Respondents were spread across grade levels with 4.4% in their first year, 20.4% in their second year, 23.0% in their third year, 38.9% in their fourth year, 9.7% in their fifth year, and 3.5% in their sixth year. Students who stopped the survey mid-way or missed a question were not included within the cluster analysis. A single participant provided the same response for each item, including reverse coded items, and was also removed from analysis. A more in-depth summary of the institutions, the survey distribution, and student responses can be found in a previous publication (Benson et al., 2016).

Evaluation of Survey Factors

Prior to performing cluster analysis, we determined the reliability of each factor, removed items that affected the reliability measure, and calculated composite scores for each participant. Due to the limited number of respondents relative to the number of items on the survey, a factor analysis was not performed. Internal consistency reliability was evaluated using Cronbach's alpha, which is often used to determine whether items are measuring the same construct within a factor (Tavakol & Dennick, 2011). Each item was systematically removed from the factor, Cronbach's alpha was recalculated and items found to lower Cronbach's alpha were reviewed to determine if there was sufficient evidence to remove them from the factor. A total of eleven items were removed from five of the factors. A summary of the factors and the final Cronbach's alpha values are in Table 2.

Table 2: Summary of final factor sizes and optimization by removal of items lowering Cronbach's Alpha. Specific items excluded are noted in Appendix A.

Factors	**Initial Number of Items within each Factor**	**Number of Items Excluded from Factor During Reliability Analysis**	**Final Cronbach's Alpha (after items were removed)**
Closed-Mindedness	7	1	0.708
Discomfort with Ambiguity	8	2	0.602
Certainty of Knowledge	8	1	0.725
Sources of Knowledge	10	4	0.630
Simplicity of Knowledge	3	0	0.494
Justification of Knowledge	9	3	0.608

One factor (Simplicity of Knowledge) was excluded from the cluster analysis due to a Cronbach's alpha below 0.6. Cronbach's alpha is dependent on the number of items in the factor (Sijtsma, 2009; Tavakol & Dennick, 2011); therefore, the Simplicity of Knowledge factor may have too few items.

Once factors were finalized, we calculated the factor scores for each participant by averaging their scores on items within the factor. Participant factor scores were then used as the inputs in the clustering algorithms.

Evaluating Outputs to Determine the Number of Clusters

There are measures to help in determining the number of clusters for the various clustering algorithms: within sum of squares, between sum of squares, and the CH index. Within sum of squares (wss) is a variability measure looking at how dispersed data points are within a single cluster and is calculated as the sum of the square of the distances between the data point and the centroid of the cluster. Between sum of squares (bss) is representative of how dispersed the clusters are within the space and is calculated as the sum of the square of the distances between the centroids of the clusters. Based on the definitions of wss and bss, we see that wss tends to zero as the number of clusters increases and bss tends to infinity. The ultimate goal of cluster analysis is to have data points within a cluster be similar and between the clusters be dissimilar, thus, we are looking to minimize wss while maximizing bss. If wss will continue to decrease and bss will continue to increase with any increase in the number of clusters, how do we know how many clusters to choose? Calinski and Harabasz introduced an index to help answer this question, which has since been coined the "CH index" (Equation 1) (1974).

$$CH\ Index = \frac{bss(k)/(k-1)}{wss(k)/(n-1)} \quad (1)$$

The CH index normalizes wss and bss relative to the number of clusters (k) and data points (n) such that the CH index is at its maximum for the best solution. Many other selection criteria exist; however, the CH index was found to be one of the most reliable (Milligan & Cooper, 1985). Note that the CH index is not defined at k = 1. As such, we cannot determine if the ideal solution is a single cluster using the CH index. This may be more concerning if the CH index was the only aspect of the data we used to determine the ideal number of clusters; however, with redundancy in our analysis we can ensure that this caveat does not negatively impact our results.

Hierarchical Clustering Methods

The first step to performing a hierarchical cluster analysis is to calculate the proximity measure. In both Ward's and Complete Link, we used Euclidean distance (Equation 2) to calculate the proximity of data points relative to each other.

$$d_{ij} = \left[\sum_k \left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}} \quad (2)$$

After the distance matrix is calculated, we used it as the input in the clustering algorithm. In the statistical language R (Team, 2016), the same function (hclust) can be used for many commonly used hierarchical clustering algorithms, including Ward's, Complete Link, and othersThe output of hierarchical cluster analysis is a dendrogram (Figure 1), or tree, that is used to visualize the merging of the data into a single cluster. On one axis are the data points within the set and on the other is a normalized distance between cluster solutions. The dendrogram allows the user to visualize every merge the clustering algorithm created and determine the appropriate number of clusters.

To determine the ideal number of clusters requires a strong command of the underlying constructs being analyzed within the data, and the assumptions made by each clustering algorithm used. Criteria to determine the number of clusters include the height of the cluster branches in the dendrogram (Figure 1A), dramatic decreases in wss (an observed 'elbow' in the plot, Figure 1B), dramatic increases in bss (Figure 1C), and the maximum CH index value (Figure 1D).

Cluster Solution of our Data Set Using Ward's

Ward's algorithm seeks to optimize wss. In Ward's, two clusters are merged that provide the smallest increase in wss until a single cluster exists. Ward's algorithm tends to create clusters of equal size but is not as sensitive to outliers in the data as much as other hierarchical approaches.

Using Ward's clustering algorithm on our survey data set, a two-cluster solution is consistently indicated (Figure 1. There is strong agreement between all four measures used to determine the number of clusters. The undulating nature of the CH index (Figure 1D) seems to indicate a small tendency towards subgroups in this cluster solution. Because the intended purpose of our cluster analysis is to select participants for follow-up interviews, these subgroups may become more useful in elucidating the differences between individuals within a larger cluster.

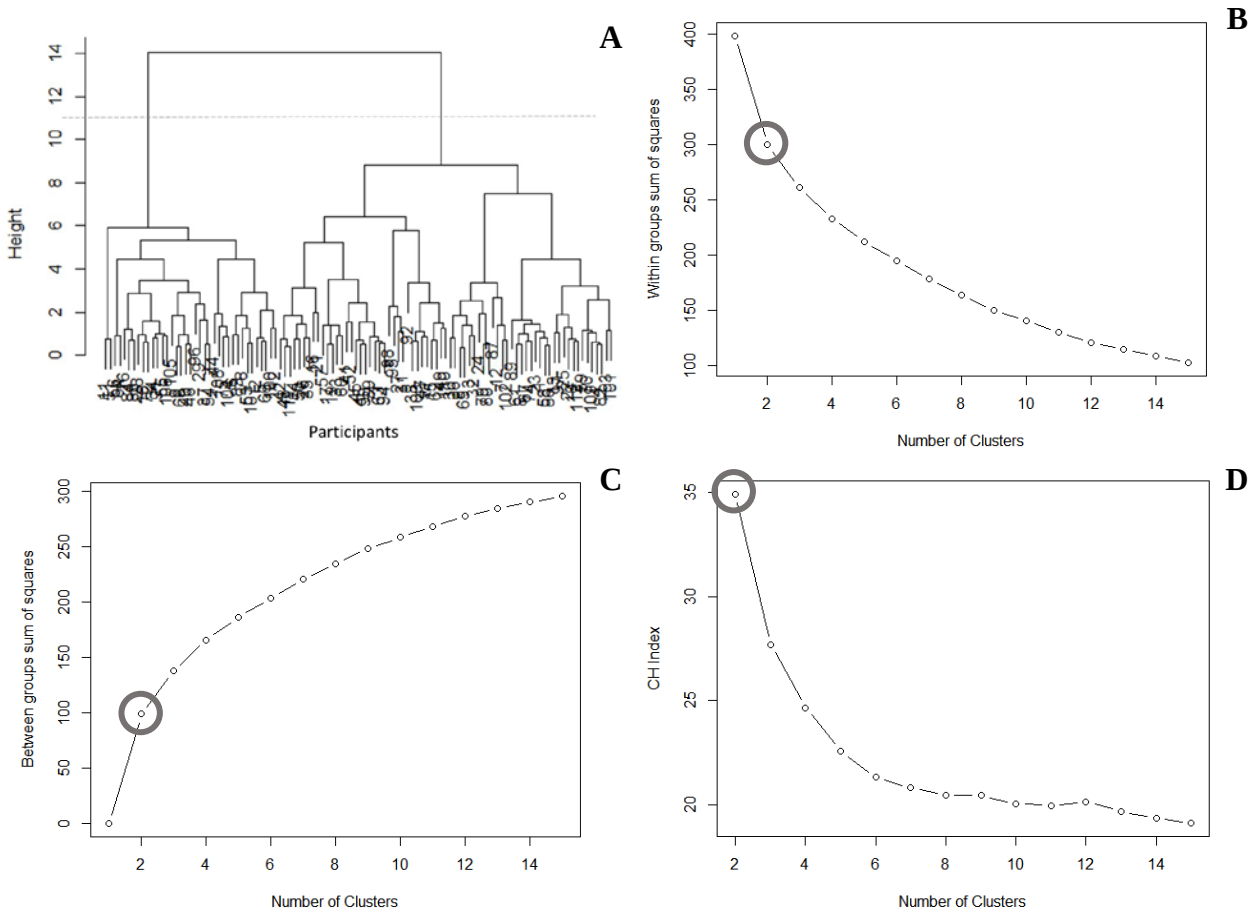

Figure 1: Plots used to determine the ideal cluster solution using Ward's algorithm and this data set is a two-cluster solution. The large height difference between one and two clusters in the dendrogram (A) as indicated by the dashed line, the "elbow" in the within sum of squares plot (B), and the "elbow" in the between sum of squares plot (C) all suggest a two-cluster solution. Additional confirmation is through the maximum CH index (D) occurs for two clusters.

The final check on the cluster solution is a plot of the data points within their assigned clusters. The further apart the clusters are in the plot, the stronger the indication of a reliable clustering solution. The two-dimensional visualization of the Ward's clustering solution (Figure 2) shows each participant and where they are located within their respective clusters for our data set. The outliers (data points 92 & 87) may be exacerbating the overlapping behavior of the clusters. As we intend to use this potential solution to help guide our participant selection for further interviews, we are less concerned with the overlapping behavior of the clusters.

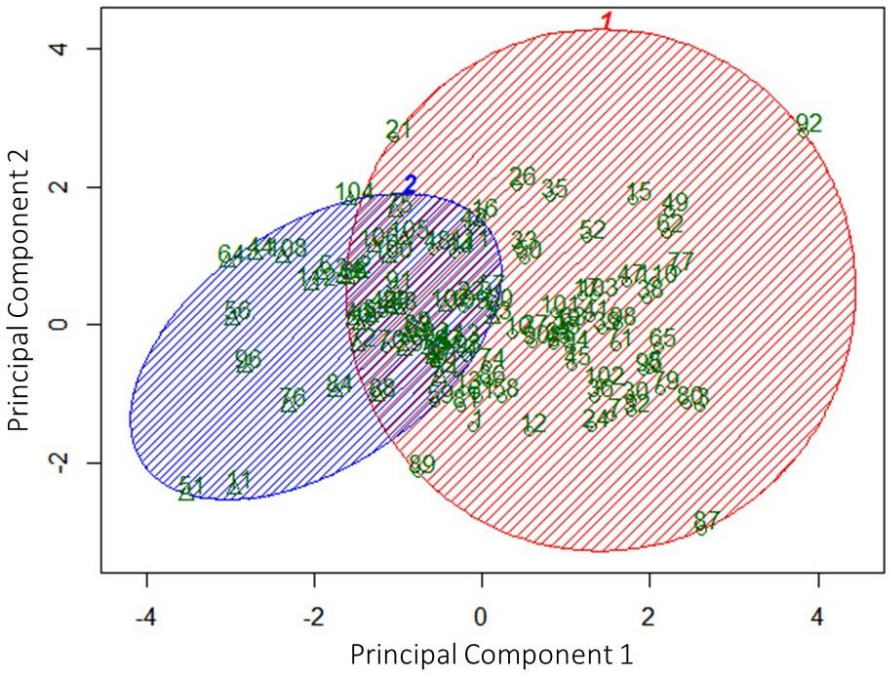

Figure 2: Two-dimensional visualization of the Ward's two cluster solution for the survey data set. The red and blue ellipses are the minimum area that incorporates all the participants within the cluster.

Cluster Solution of our Data Set Using Complete Link

The Complete Link algorithm merges clusters based on distance between points currently within the clusters. The cluster solution for our data set is similar to the Ward's solution in terms of the branch heights in the dendrogram for multiple cluster solutions (Figure 3A), the elbows observed in wss (Figure 3B) and bss (Figure 3C), as well as the maximum behavior of the CH index plot (Figure 3D) all strongly indicate a two-cluster solution. In the dendrogram we also see participant 92 remaining in their own cluster until the very end. This would indicate that this participant has very different views relative to the rest of the group and thus could provide an especially unique viewpoint in an interview.

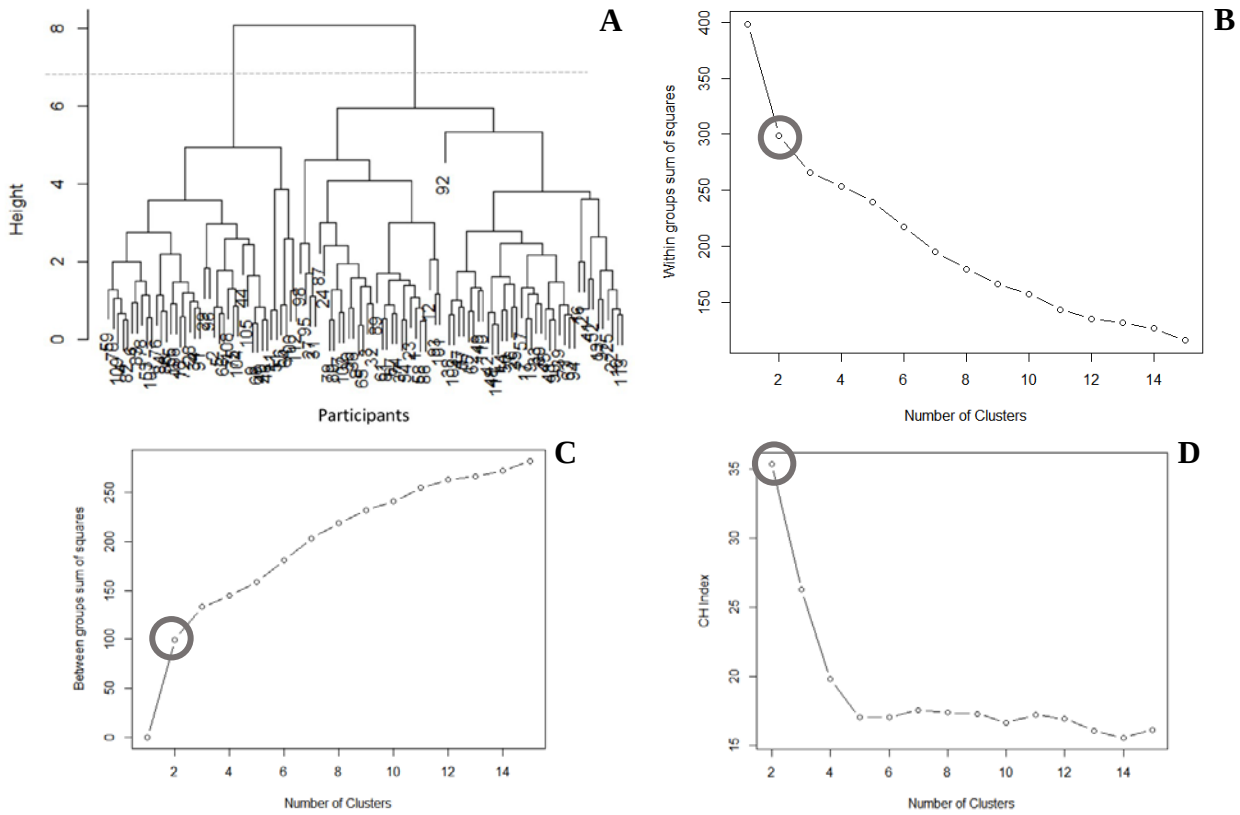

Figure 3: Plots used to determine the ideal cluster solution using the Complete Link algorithm and this data set is a two-cluster solution. The large height difference between one and two clusters in the dendrogram (A) as indicated by the dashed line, the "elbow" in the within sum of squares plot (B), and the "elbow" in the between sum of squares plot (C) all suggest a two-cluster solution. Additional confirmation is through the maximum CH index (D) occurs for two clusters.

The Complete Link cluster plot (Figure 4) is similar to the Ward's cluster plot. Again, the outliers are impacting the size of the ellipse of cluster 1. The differences between cluster assignment between the Ward's and Complete Link algorithms is most likely occurring in the overlap of the two clusters.

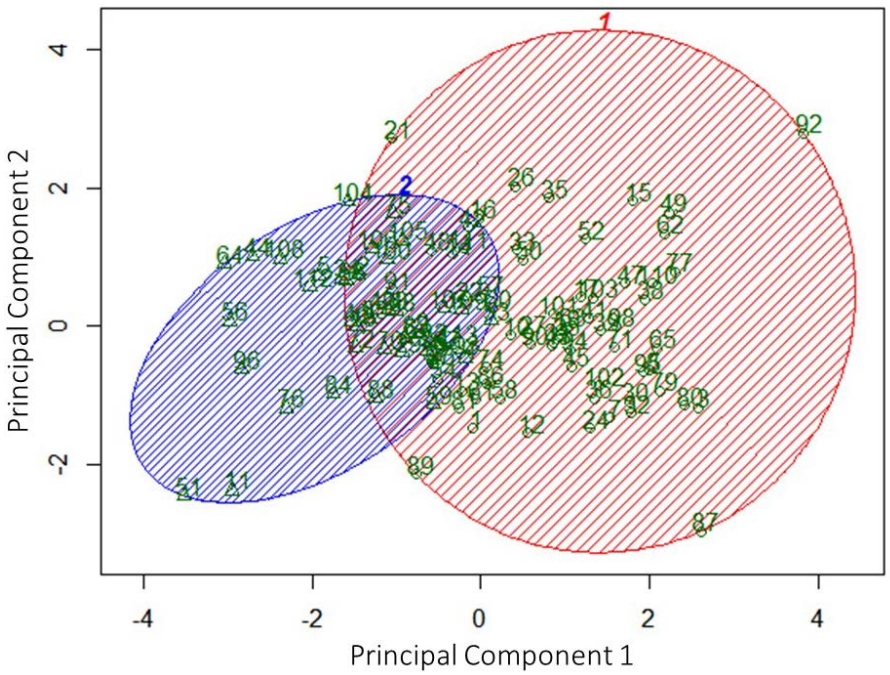

Figure 4: Two-dimensional visualization of the Complete Link two-cluster solution for the survey data set. The red and blue ellipses are the minimum area that incorporates all the participants within the cluster.

Cluster Solution of our Data Set Using K-means

A partitioning algorithm requires the user to input the number of clusters prior to optimization, thus, it is important to either have strong theoretical support for the number of clusters selected or run the algorithm for multiple cluster solutions. Because of the previous solutions with hierarchical algorithms, we anticipate the solution will indicate two clusters; however, we still need to confirm this prediction. Similar to our approach with the hierarchical algorithms, wss, bss, the CH index will be plotted for cluster solutions up to 15 clusters and then the number of clusters will be determined from the plots of our data set. As this is a partitioning algorithm with many cases switching cluster assignments on the same iteration, a dendrogram is not created. Wss, bss, and the CH index (Figure 5) all indicate that an appropriate number of clusters for this data set is two. K-means also seems to provide the best visual solution for our data set, seen in the cluster plot (Figure 6), as there is the most amount of separation of the clusters.

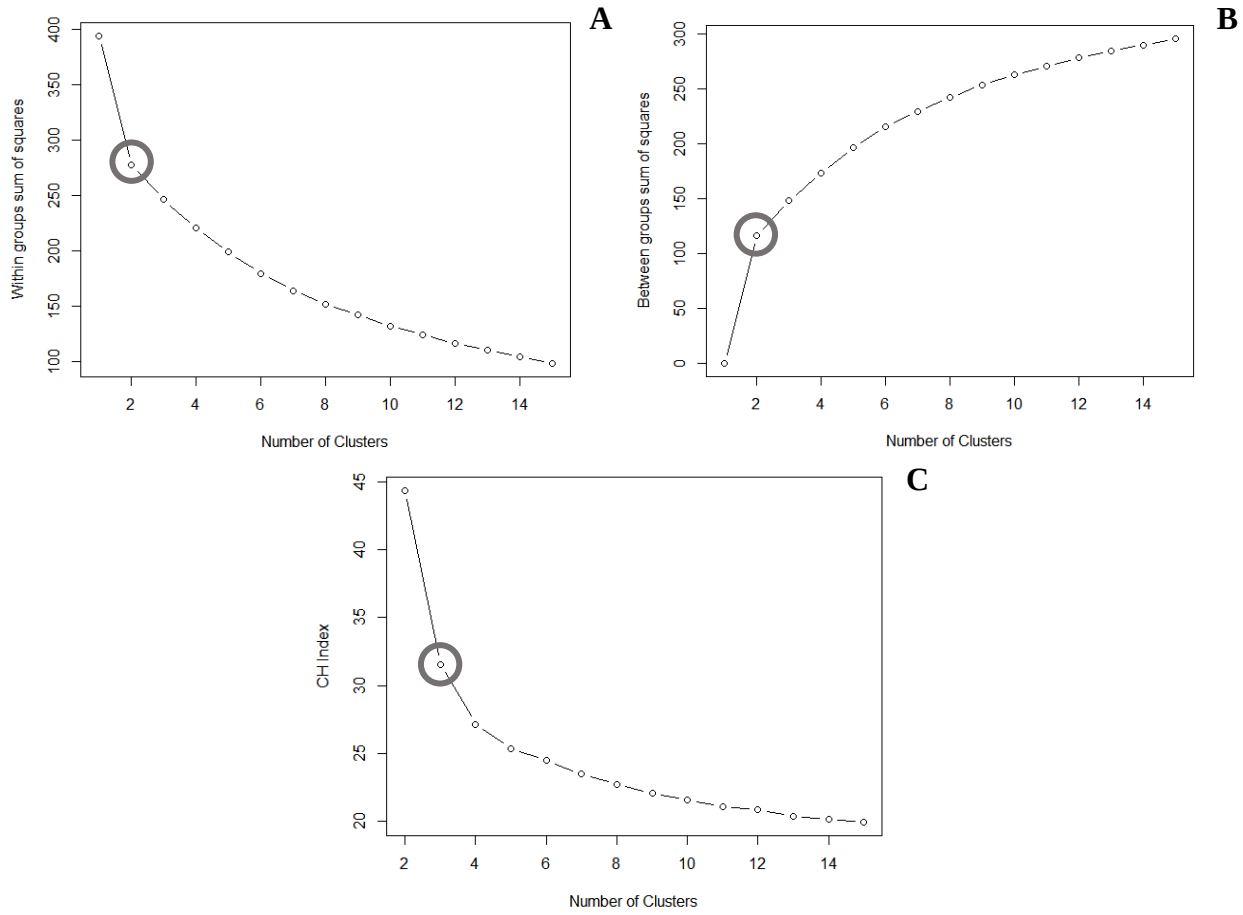

Figure 5: Plots used to determine the ideal cluster solution using the k-means algorithm and this data set is a two-cluster solution. The “elbow” in the within sum of squares plot (A), and the “elbow” in the between sum of squares plot (B) all suggest a two-cluster solution. Additional confirmation is through the maximum CH index (C) occurs for two clusters.

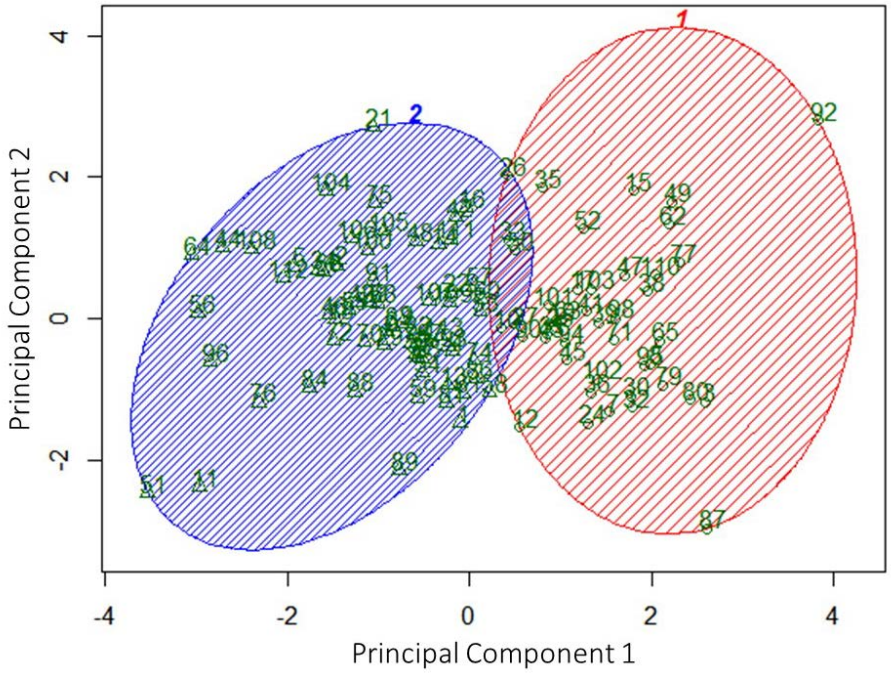

Figure 6: Two-dimensional visualization of the k-means two cluster solution for the survey data set. The red and blue ellipses are the minimum area that incorporates all the participants within the cluster.

Comparing Cluster Solutions

As described above, cluster solutions indicate that a two-cluster solution is the ideal number of clusters for our data set. A summary of the construct means for each cluster and algorithm combination is shown in Table 3.

Table 3: Summary of the average scores for the clusters on each factor used during cluster analysis. The total number of participants in each cluster is also provided at the bottom of the table.

	Ward's		**Complete Link**		**k-means**	
Factor	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Closed-Mindedness	2.75 ± 0.75	3.68 ± 0.73	2.69 ± 0.71	3.72 ± 0.71	2.53 ± 0.74	3.43 ± 0.75
Discomfort with Ambiguity	4.84 ± 0.92	4.74 ± 0.76	4.85 ± 0.94	4.74 ± 0.72	4.71 ± 0.93	4.87 ± 0.82
Certainty of Knowledge	2.04 ± 0.59	3.27 ± 0.72	2.06 ± 0.63	3.14 ± 0.80	1.80 ± 0.51	2.92 ± 0.76
Sources of Knowledge	3.81 ± 0.80	4.48 ± 0.74	3.76 ± 0.80	4.50 ± 0.70	3.42 ± 0.74	4.46 ± 0.61
Justification of Knowledge	3.24 ± 0.75	4.33 ± 0.60	3.21 ± 0.75	4.30 ± 0.59	2.93 ± 0.66	4.08 ± 0.67
Number of Participants (n)	71	37	68	40	44	64

We compared the sets of clusters to each other to determine if there was a statistically significant difference between the cluster means. First, we performed a multivariate analysis of variance (MANOVA) to determine if there were differences between cluster groups when considering all five factors at the same time and the interaction between the clustering technique with the assigned cluster. Significant differences were observed with a p-value $p < 2e\text{-}16$ for the interaction test indicating that some differences existed, but did not specifically indicate the differences.

To determine where the differences occurred, we conducted individual analysis of variance (ANOVA) tests on each factor to determine which factor(s) were significantly different between the sets of clusters. The cluster means were significantly different on four of the five factors: Closed-Mindedness ($p < 2e\text{-}16$), Certainty of Engineering Knowledge ($p < 2e\text{-}16$), Sources of Engineering Knowledge ($p < 2e\text{-}16$), and Justification of Engineering Knowledge ($p < 2e\text{-}16$). The only factor that was not significant was Discomfort with Ambiguity ($p = 0.964$). The results of the ANOVA tests indicate that at least one of the six clusters (cluster 1 or 2 from any of the three techniques) is different than the other six clusters for the factors that showed significant differences.

To determine which clusters differed and in what ways, we performed a Tukey HSD test for each factor that indicated significance in the ANOVA. The only factor that we did not conduct a Tukey HSD test on was the Discomfort with Ambiguity factor, because no significant differences between any of the clusters were identified by the ANOVA test. A Tukey HSD test is a pairwise test to evaluate the differences between all combinations of clustering technique and cluster assignment.

Results of the Tukey HSD tests indicated that the significant differences observed were between clusters within a single technique (i.e. cluster assignment) and not between clustering techniques. For example, in the Certainty of Knowledge factor, the significant differences only occur between clusters within a single technique: Cluster 1 from the Ward's algorithm aligned with Cluster 1 of both the Complete Link and the k-means algorithms. Cluster 2 from the Ward's algorithms aligned with Cluster 2 for both the Complete Link and k-means algorithms. Differences between clusters are summarized in Table 4 and Figure 7. A full summary of the Tukey HSD results is included in Appendix B.

Table 4: Summary of Tukey HSD results. Scores are based on a 7-point scale where 1 is strongly disagree, 4 is neutral, and 7 is strongly agree. An asterisk (*) is placed on the averages that differ within a single technique (i.e. Cluster 1 compared to Cluster 2 in the k-means column).

	Ward's		**Complete Link**		**k-means**	
Factor	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Closed-Mindedness	2.75 ± 0.75*	3.68 ± 0.73*	2.69 ± 0.71*	3.72 ± 0.71*	2.53 ± 0.74*	3.43 ± 0.75*
Discomfort w. Ambiguity	4.84 ± 0.92	4.74 ± 0.76	4.85 ± 0.94	4.74 ± 0.72	4.71 ± 0.93	4.87 ± 0.82
Certainty of Knowledge	2.04 ± 0.59*	3.27 ± 0.72*	2.06 ± 0.63*	3.14 ± 0.80*	1.80 ± 0.51*	2.92 ± 0.76*
Sources of Knowledge	3.81 ± 0.80*	4.48 ± 0.74*	3.76 ± 0.80*	4.50 ± 0.70*	3.42 ± 0.74*	4.46 ± 0.61*
Justification of Knowledge	3.24 ± 0.75*	4.33 ± 0.60*	3.21 ± 0.75*	4.30 ± 0.59*	2.93 ± 0.66*	4.08 ± 0.67*
Number of Participants (n)	71	37	68	40	44	64

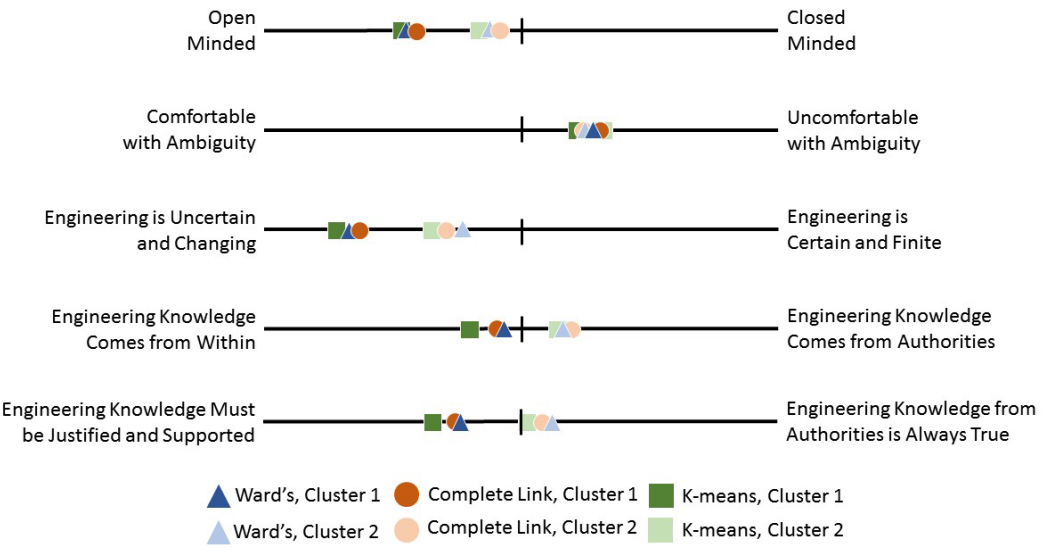

Figure 7: Visual representation of the cluster means on each factor. Cluster means are plotted on each factor scale. Ward's clusters are indicated by the blue triangles, Complete Link by the orange circles, and k-means by green squares. The dark shapes represent the mean for Cluster 1 and the light shapes represent the mean of Cluster 2. Neutral point was indicated with a black dash and descriptors of the ends are provided.

Conclusions

With our data set, we see agreement between three clustering algorithms in the number of clusters and overall cluster means. After reviewing the within-cluster sum of squares, between-cluster sum of squares, CH index, and cluster solution plots, we determined the most reliable clustering algorithm for this data set to be k-means. Using multiple clustering algorithms ensured we had strong agreement in the number and shape of the clusters. Additionally, we identified outlier participants that may be interesting to recruit for our follow-up interviews.

Choosing an appropriate clustering algorithm to sort data into homogenous groups depends on the nature of the data (number of data points, the spread of the data, etc.), and whether there is underlying theory that would help predict the number of clusters. Algorithms differ in terms of how they group the data, their sensitivity to outliers, and computational efficiency. Hierarchical methods are typically chosen when there is not strong underlying theory and small sample sizes. It runs through the range from n to 1 clusters. Ward's tries to minimize within-group sum of squares and maximize between-group sum of squares. Complete Link algorithm can be sensitive to outliers because it evaluates items that are furthest apart. K-means is the most commonly used partitioning method and is computationally efficient; however, the number of clusters is assumed based on *a priori* knowledge (underlying theory, the nature of the data, other clustering solutions, etc.). Based on the data presented above, the clustering algorithm that seems most ideal for the data set analyzed in this study is k-means. The k-means solution had clusters that were visually furthest apart and cluster means that are sufficiently similar to the other clustering solutions. Wss, bss, and the CH index strongly indicated across all three solutions that two clusters was the ideal number of clusters for this data set.

Limitations and Future Work

Although differences in the clustering solutions and the factor means are observed, we do see sufficient convergence of the results at the completion of this portion of the study. Due to the foundational assumptions of all three clustering algorithms, we may not be able to determine the natural tendency of our population of interest (in this case, undergraduate mechanical or biomedical engineers with research experience) from these results; however, we are able to group individuals with similar scores together. Since we are utilizing the results of the cluster analysis for participant selection only, grouping participants is sufficient, even if it might not be the natural tendency of the data. If we were to use the cluster analysis results to create develop a new theoretical framework on epistemic beliefs, a more rigorous analysis of the clustering results additional measures of the participants relative to the cluster data would be required.

Now that we have selected our clustering algorithm and have determined the cluster solution, we will be able to select participants for interviews based on their location within the cluster. In this mixed-methods study, we will use the cluster assignment to ensure that we select participants with a range of epistemic beliefs and need for cognitive closure. This approach will help use ensure that we are interviewing participants with a wide-range of beliefs to inform the development of our grounded theory.

Implications for Practice

Accurately running and analyzing the results of cluster analysis can be a difficult if you are not familiar with without a thorough understanding of the foundational assumptions made when applying an algorithm to your data set. We suggest you take the following a series of steps when deciding which cluster analysis method to perform your ownapply, and when performing a cluster analysis as described below and depicted in Figure 8.

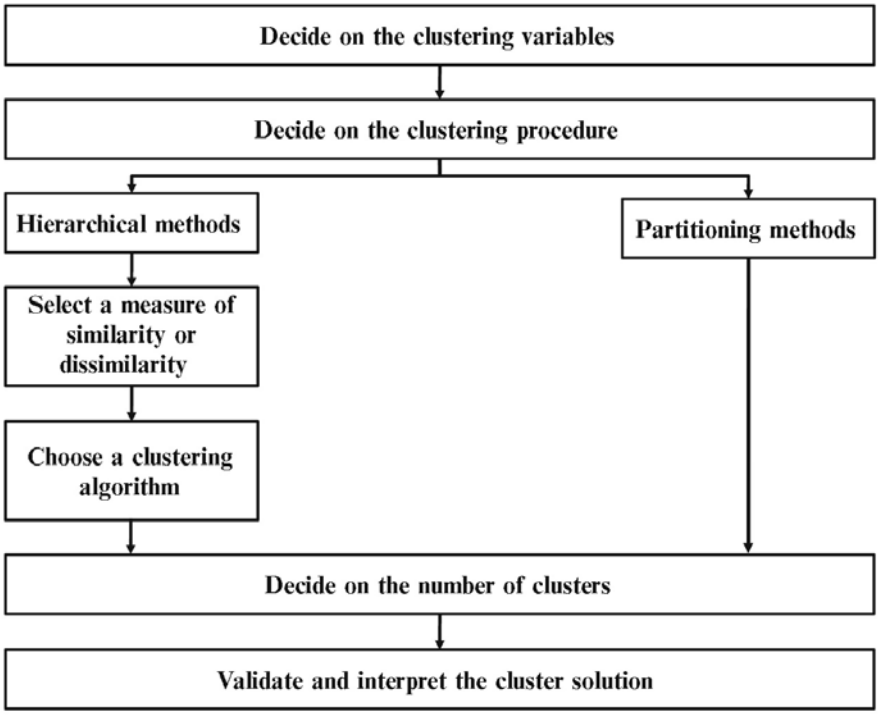

Figure 8: Flow chart for performing a cluster analysis. This chart was modified based on earlier work (Mooi & Sarstedt, 2011) to fit the context of this paper.

1. Determine the variables that you will be using for your cluster analysis.
    a. These should be related to your research questions and/or grounded in the theoretical lenses you are using in your study.
    b. If your variables are constructs or factors from survey items, check for internal consistency using a standard measure like Cronbach's alpha.
2. Determine the clustering procedure(s) to use for your cluster analysis
    a. Review the data set and the theoretical grounding for the research
        i. If the data set is small or you do not have theoretical grounding for knowing the number of clusters, rely mostly on hierarchical clustering procedures
        ii. If the data set is large or you have theoretical grounding for knowing the number of clusters *a priori*, rely on partitioning algorithms
3. Determine the similarity/dissimilarity measure (for hierarchical algorithms only)
4. Determine the algorithm(s) to use based on data behavior (for hierarchical algorithms only)
    a. If data clumps in an oblong way, select Single Link.
    b. If data clumps in a circular way, select other hierarchical clustering algorithms
5. Determine the number of clusters
    a. Evaluate dendogramsdendrograms (hierarchal algorithms only) and other plot (wss, bss, CH index, etc.) behavior to help determine the number of clusters
6. Validate and interpret the solution(s) provided.
    a. If using multiple solutions, use MANOVA and ANOVA testing to determine solution convergence.

Acknowledgements

The authors would like to thank Ms. Paran Norton and Mr. William Bridges for their assistance in the statistical comparison of the cluster solutions. This research is supported by the National Science Foundation [Award #s 1531607, 1531641]. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation.

References

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. http://doi.org/10.1007/s11423-012-9235-8

Benson, L. C., Kennedy, M. S., Katherine, M., Carolina, S., Faber, C. J., Kajfez, R. L., … Vargas, P. M. D. (2016). WIP : Understanding Undergraduate Engineering Researchers and How They Learn, 0–4.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. http://doi.org/10.1080/03610927408827101

Chan, J. Y. K., & Bauer, C. F. (2014). Identifying At-Risk Students in General Chemistry via Cluster Analysis of Affective Characteristics. *Journal of Chemical Education*. Easton: Division of Chemical Education, Inc and ACS Publications Division of the American Chemical Society.

Faber, C., Vargas, P., & Benson, L. (n.d.). Measuring Engineering Epistemic Beliefs in Undergraduate Engineering Students.

Ferguson, L. E., & Braten, I. (2013). Student profiles of knowledge and epistemic beliefs: Changes and relations to multiple-text comprehension. *Learning and Instruction*, *25*, 49–61. http://doi.org/10.1016/j.learninstruc.2012.11.003

Galloway, K. R., & Bretz, S. L. (2015a). Measuring meaningful learning in the undergraduate chemistry laboratory: a national, cross-sectional study. *Journal of Chemical Education*. Easton: American Chemical Society Division of Chemical Education.

Galloway, K. R., & Bretz, S. L. (2015b). Using cluster analysis to characterize meaningful learning in a first-year university chemistry laboratory course. *Chem. Educ. Res. Pract*. http://doi.org/10.1039/c5rp00077g

Greene, J. A., Torney-Purta, J., & Azevedo, R. (2010). Empirical evidence regarding relations among a model of epistemic and ontological cognition, academic performance, and educational level. *Journal of Educational Psychology*, *102*(1), 234–255. http://doi.org/10.1037/a0017998

Jiang, I. G., Yeh, L. C., Hung, W. L., & Yang, M. S. (2006). Data analysis on the extrasolar planets using robust clustering. *Monthly Notices of the Royal Astronomical Society*, *370*(3), 1379–1392. http://doi.org/10.1111/j.1365-2966.2006.10580.x

Karabenick, S. A. (2003). Seeking help in large college classes: A person-centered approach. *Contemporary Educational Psychology*, *28*(1), 37–58. http://doi.org/10.1016/S0361-476X(02)00012-7

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY: Wiley.

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing Multivariate Data* (1st ed.). Pacific Grove, CA: Books/Cole - Thomson Learning.

Lorentz, H., Hilmola, O. P., Malmsten, J., & Srai, J. S. (2016). Cluster analysis application for understanding SME manufacturing strategies. *Expert Systems with Applications*, *66*, 176–188. http://doi.org/10.1016/j.eswa.2016.09.016

Maccuish, J. D., & Maccuish, N. E. (2014). Chemoinformatics applications of cluster analysis. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(1), 34–48. http://doi.org/10.1002/wcms.1152

Milligan, G. W., & Cooper, M. C. (1985). An Examination of Procedures for Determining the

Number of Clusters in a Data Set. *Psychometrika, 50*(2), 159–179.
Mooi, E., & Sarstedt, M. (2011). Cluster Analysis. In *A Concise Guide to Market Research* (pp. 237–284). http://doi.org/10.1007/978-3-642-12541-6
Raker, J. R., Reisner, B. A., Smith, S. R., Stewart, J. L., Crane, J. L., Pesterfield, L., & Sobel, S. G. (2015). Foundation Coursework in Undergraduate Inorganic Chemistry: Results from a National Survey of Inorganic Chemistry Faculty. *Journal of Chemical Education*. Easton: Division of Chemical Education, Inc and ACS Publications Division of the American Chemical Society.
Rencher, A. C. (2002). *Methods of multivariate analysis* (Second). New York, NY: John Wiley & Sons.
Shell, D. F., & Soh, L. K. (2013). Profiles of Motivated Self-Regulation in College Computer Science Courses: Differences in Major versus Required Non-Major Courses. *Journal of Science Education and Technology, 22*(6), 899–913. http://doi.org/10.1007/s10956-013-9437-9
Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach. *Psychometrika*, 107–120. http://doi.org/10.1007/s11336-008-9101-0
Stewart, J., Miller, M., Audo, C., & Stewart, G. (2012). Using Cluster Analysis to Identify Patterns in Students' Responses to Contextually Different Conceptual Problems . *Physical Review Special Topics - Physics Education Research* . American Physical Society . http://doi.org/10.1103/PhysRevSTPER.8.020112
Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. http://doi.org/10.5116/ijme.4dfb.8dfd
Team, R. C. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062.
Yu, J. H., & Strobel, J. (2012). A First Step in the Instrument Development of Engineering-related Beliefs Questionnaire. *Proceedings of the 2012 American Society for Engineering Education Annual Conference and Exposition*.

Appendix A: Final item wording on Phase I survey

An (RC) indicates items that were reverse-coded and items with a strikethrough were excluded from cluster analysis.

Item	Testing	Factor	Source
~~Classroom engineering problems have only one right numerical answer.~~	Engineering Epistemic Beliefs	Certainty of Knowledge	Yu & Strobel 2012
Engineering knowledge cannot be subject to change with new observations by engineering students.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Theories in engineering cannot be argued or changed.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Engineering problems outside the classroom have only one right numerical answer.	Engineering Epistemic Beliefs		Yu & Strobel 2012
There is one universal engineering method.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Engineering knowledge cannot be subject to change with new observations by individuals.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Engineering knowledge is all factual, and there are no opinions.	Engineering Epistemic Beliefs		Greene et al. 2010
Engineering knowledge should be accepted as an unquestionable truth.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Engineering students learn when a teacher transmits his or her knowledge to them.	Engineering Epistemic Beliefs	Sources of Knowledge	Yu & Strobel 2012
In an engineering class, if your personal experience conflicts with the "big ideas" in a textbook, the textbook is probably right.	Engineering Epistemic Beliefs		Yu & Strobel 2012
You can trust the information you find in engineering textbooks.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Engineering knowledge is created only by an expert.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Reading engineering textbooks written by experts is the best way to learn engineering.	Engineering Epistemic Beliefs		Yu & Strobel 2012
Traditional engineering ideas should be considered over new ideas.	Engineering Epistemic Beliefs		Yu & Strobel 2012

~~First-hand experience is the best way of knowing something in engineering. (RC)~~	Engineering Epistemic Beliefs	Sources of Knowledge	Yu & Strobel 2012
Engineering knowledge is created only from logical thinking.	Engineering Epistemic Beliefs		Yu & Strobel 2012
The best way to develop engineering knowledge is from an expert's teachings.	Engineering Epistemic Beliefs		Yu & Strobel 2012
~~New engineering knowledge is produced as a result of experimentation.~~	Engineering Epistemic Beliefs		Yu & Strobel 2012
~~Engineers can solve engineering problems by just following a step-by-step procedure.~~	Engineering Epistemic Beliefs	Simplicity of Knowledge	Yu & Strobel 2012
~~To know engineering well, you need to memorize what you are taught.~~	Engineering Epistemic Beliefs		Greene et al. 2010
~~Engineering knowledge is an accumulation of facts.~~	Engineering Epistemic Beliefs		Yu & Strobel 2012
In engineering, what's a fact depends upon a person's point of view. (RC)	Engineering Epistemic Beliefs	Justification of Knowledge	Greene et al. 2010
To be able to trust knowledge claims in engineering, I have to check various sources. (RC)	Engineering Epistemic Beliefs		Ferguson & Braten 2013
In engineering classes, everyone's knowledge can be different because there is no one absolutely right answer. (RC)	Engineering Epistemic Beliefs		Greene et al. 2010
If an engineer says something is a fact, I believe it.	Engineering Epistemic Beliefs		Greene et al. 2010
~~Just one source is never enough to decide what is right in engineering. (RC)~~	Engineering Epistemic Beliefs		Greene et al. 2010
I believe everything I learn in my engineering classes.	Engineering Epistemic Beliefs		Greene et al. 2010
~~To detect incorrect claims in texts about engineering, it is important to check several information sources. (RC)~~	Engineering Epistemic Beliefs		Ferguson & Braten 2013
If an engineering teacher says something is a fact, I believe it.	Engineering Epistemic Beliefs		Greene et al. 2010
In the field of engineering, everyone's knowledge can be different because there is no one absolutely right answer. (RC)	Engineering Epistemic Beliefs		Greene et al. 2010

Even after I've made up my mind about something, I am always eager to consider a different opinion. (RC)	Need for Cognitive Closure	Closed-Mindedness	Webster & Kruglanski, 1994
I prefer interacting with people whose opinions are very different from my own. (RC)	Need for Cognitive Closure		Webster & Kruglanski, 1994
I dislike questions which could be answered in many different ways.	Need for Cognitive Closure		Webster & Kruglanski, 1994
I do not usually consult many different opinions before forming my own view.	Need for Cognitive Closure		Webster & Kruglanski, 1994
~~I always see many possible solutions to problems I face. (RC)~~	Need for Cognitive Closure		Webster & Kruglanski, 1994
When considering most conflict situations, I can usually see how both sides could be right. (RC)	Need for Cognitive Closure		Webster & Kruglanski, 1994
When thinking about a problem, I consider as many different opinions on the issue as possible. (RC)	Need for Cognitive Closure		Webster & Kruglanski, 1994
It's annoying to listen to someone who cannot seem to make up his or her mind.	Need for Cognitive Closure	Discomfort with Ambiguity	Webster & Kruglanski, 1994
I don't like situations that are uncertain.	Need for Cognitive Closure		Webster & Kruglanski, 1994
I feel uncomfortable when someone's meaning or intention is unclear to me.	Need for Cognitive Closure		Webster & Kruglanski, 1994
When I am confused about an important issue, I feel very upset.	Need for Cognitive Closure		Webster & Kruglanski, 1994
I'd rather know bad news than stay in a state of uncertainty.	Need for Cognitive Closure		Webster & Kruglanski, 1994
I feel uncomfortable when I don't understand the reason why an event occurred in my life.	Need for Cognitive Closure		Webster & Kruglanski, 1994
~~In most social conflicts, I can easily see which side is right and which is wrong.~~	Need for Cognitive Closure		Webster & Kruglanski, 1994
I like to know what people are thinking all the time.	Need for Cognitive Closure		Webster & Kruglanski, 1994