A Cluster Analysis of Domain Specific Self-Efficacy in College Students

A Research Project

Submitted to the Graduate School

In Partial Fulfillment of the Requirements

For the Degree

Master of Science

By

Raymond Caridi

Dr. Holmes Finch – Advisor

Ball State University

Muncie, Indiana

May, 2020

The psychological concept of one’s confidence in their ability to perform tasks or to overcome obstacles is known as one’s self-efficacy. Self-efficacy is the belief in one’s own capability to perform a task (Gist, 1987). Self-efficacy has been found to be positively associated with overall well-being, which refers to an individual’s sense of how satisfied they are with the trajectory of their life. Tong and Song (2004) examined this relationship between self-efficacy and well-being in college students in China. Tong and Song (2004) defined self-efficacy as a person’s optimistic self-belief to cope with a variety of difficult demands in life, and they defined subjective well-being as a person’s general affect and overall life satisfaction. Students who displayed higher levels of self-efficacy also reported having greater levels of overall general well-being (Tong & Song, 2004).

Alongside Tong and Song’s study, there are other studies conducted previously which support the claim that self-efficacy is correlated to overall well-being. However, a limitation of these studies is that they only examined general self-efficacy – individuals’ overall sense of how competent they were in their lives - and its relationship to subjective well-being. An individual may not feel competent in every one of their roles; rather, individuals often feel varying levels of competence and confidence in different areas of life. A person may have particularly high self-efficacy in one area of life, and low self-efficacy in another area. This focused sense of self-efficacy is referred to as domain self-efficacy. Domain self-efficacy is defined as one’s belief in their abilities to perform a task in only one specific area or domain, such as school or work for example. Lent, Brown, and Gore Jr. (1997) attempted to differentiate domain self-efficacy from general self-efficacy. In their study, 205 university students answered questions which revealed their personal levels of academic self-concept, global academic self-efficacy, and domain-specific mathematics self-efficacy. Their results showed that students reported having latent dimensions of self-perception and different levels of self-efficacy and self-concept in different areas of life which were distinguishable from one another (Lent et al., 1997). Thus, people can have different levels of confidence in different areas of life - such as in school, at work, at home, or in hobbies. Many studies examine self-efficacy in a general sense, where people are assumed to have a general sense of overall self-efficacy. While this is true, one's self-efficacy is complex, and can be broken down into different parts within different domains.

When examining research conducted about complex topics that include multiple variables, such as the different domains of self-efficacy, for example, it is necessary to use the correct methods. One such method that can be used would be a cluster analysis. There are several different types of cluster analysis, which all have different purposes and can be used to analyze different situations. Two popular approaches to cluster analysis are Hierarchical methods and non-hierarchical methods. The most popular form of the hierarchical method is the agglomerative method, where each subject starts out in their own specific cluster. The two most similar clusters in the sample are then combined into one, new cluster. This is done repeatedly until the optimal number of clusters is achieved from the data. This hierarchical agglomerative cluster analysis is used in a study conducted by Paul D. Loprinzi and Jerome F. Walker in 2016 (Loprinzi & Walker, 2016). The authors of this study wanted to examine the longitudinal effects of changes in physical activity on the likelihood to quit smoking in a nation sample of daily smokers, aged 16-24 years old. They surveyed the participants over the course of three years about their demographic information, their levels of physical activity, and how often the continued to smoke, if at all. In order to best measure physical activity in participants, the researchers of this study utilized Ward's method of hierarchical agglomerative cluster analysis to define a necessary number of physical activity clusters across the three time periods where participants were surveyed (Loprinzi & Walker, 2016). Ward's method of hierarchical agglomerative clustering is used in this study, which is a method where all possible pairs of clusters are combined and the sum of the squared distances in each cluster is calculated (Loprinzi & Walker, 2016; Cornish, 2007). It is important for this method of cluster analysis, and other methods of hierarchical agglomerative cluster analysis as well, what "distance" actually is. In cluster analysis, the distance is the dissimilarity between each pair of observations (Kassambara, 2018). There are several different equations a researcher can use to measure distance in a cluster analysis, and the method that is chosen is very important. One of the most common distance measuring methods is called the Euclidean distance (Kassambara, 2018). The Euclidean distance states that in general, if you have p variables (X_p) measured based on a sample of n participants, then the observed data for subject i can be denoted by X_{ip} and the observed data for subject j can be denoted by X_{jp} (Kassambara, 2018; Cornish, 2007). The equation for the Euclidean distance between the two subjects is written as:

$$d_{ij} = \sqrt{}((X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \ldots + (X_{ip} - X_{jp})^2)$$

In other situations or studies, a different distance measure may be used, such as the Pearson correlation distance. This measure is used to measure the degree of a linear relationship between two profiles, and finds objects to be similar when two or more of them are highly correlated with one another (Kassambara, 2018). The equation for the Pearson correlation distance is:

$$d_{cor}(x,y) = 1 - \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^{2\Sigma}}(y_i - \bar{y})\hat{}2}$$

One other popular distance measure is the Manhattan distance measure. This method is very similar to the Euclean distance method, but instead of squaring the sum to get a positive number and then taking the square root of that number, the Manhattan distance simply takes the absolute value of the sum (Kassambara, 2018).

$$d_{ij} = |(X_{i1} – X_{j1}) + (X_{i2} – X_{j2}) + \ldots + (X_{ip} – X_{jp})|$$

When using the agglomerative hierarchical method, there are several other, different clustering methods within this topic one can use. There is the nearest neighbor method, where the distance between two clusters is measured by the distance between the two closest clusters. On the other end, there is also the furthest neighbor method, where the distance between two clusters is measured by the distance between the two furthest clusters. Additionally, there is the average linkage method where the distance between clusters is measured by calculation the average distance between all pairs of subjects in two clusters, as well as the centroid method, where the mean value for each variable, also known as the centroid, is calculated and the distance between these centroids is used.

Non-hierarchical clustering methods, often referred to as k-means clustering methods, are generally used when there is a large dataset due to the fact that non-hierarchical clustering allows for subjects to move from one cluster to another if necessary, which is not possible in hierarchical clustering methods (Cornish, 2007). K-means methods are run differently than hierarchical agglomerative methods, where the researcher specifies the desired number of clusters in advance, and the best solution is chosen by the researcher. The steps for this method include choosing initial cluster centre, which is the center of a point of a circle or cluster and is found by dividing each subject into a cluster of one and defining its centre as the value of the variables for that subject, then assigning each subject to its nearest cluster in terms of distance to the centroid, which is defined as the mean value for each variable (Cornish, 2007). Next, find the formed centroids, calculate the distance again from each subject to each centroid and move any data points that are not in the cluster that they are closest to (Cornish, 2007). Another important part of preforming a cluster analysis is identifying the type of data present in a study, along with ways to measure the distance of the variables present. Researchers have to identify if the data is interval data, categorical data, binary data, or ordinal data.

Another popular algorithm for clustering, besides hierarchical and k-means, is fuzzy clustering. Fuzzy clustering, sometimes referred to as soft clustering, differs from k-means clustering and other forms of clustering, referred to as hard clustering, because it allows data points to belong to more than one cluster. When using hard clustering techniques, each data point belongs to a specific cluster, and that cluster alone, but when we utilize fuzzy clustering techniques, we are capable of blurring the line between clusters, which allows data points to belong to more than one cluster (Suganya & Shanthi, 2012; Glen, 2016). When a researcher used a fuzzy cluster technique, the have the option to use either a classical fuzzy algorithm or a shape-based fuzzy algorithm. When using a classical algorithm approach, the researcher can follow the fuzzy C-means algorithm, or FCM, which is a common classical algorithm. The use of this algorithm gives each data point a membership grade between 0 and 1, where a 0 means that a data point is the furthest it can possibly be from a cluster's center, while a 1 means that a data point is as close as possible to a cluster's center (Suganya & Shanthi, 2012; Glen, 2016). Fuzzy C-means algorithms usually assume spherical clusters, however, the Gustafson-Kessel (GK) algorithm assumes an elliptical cluster shape, and associates a data point with a matrix and cluster (Suganya & Shanthi, 2012; Glen, 2016). Finally, the Gaussian Mixture Decomposition algorithm is similar to these other classical algorithms, but it allows the clusters to take on any shape. In addition to classical fuzzy clustering algorithms are shape-based fuzzy clustering algorithms. These include circular shaped, with clusters taking on the shape of circles, elliptical shaped, with clusters being more of an elliptical shape than a circle, and generic shaped, where clusters can be any shape (Suganya & Shanthi, 2012; Glen, 2016). This form of clustering allows there to be an alternative to hierarchical clustering techniques and k-means clustering techniques, and allows Bayesian statistical analysis to be even more open to models which fit data in a better matter.

Researchers have spent a lot of time examining cluster analysis, finding the best uses of clustering for different situations. Using cluster analysis can help to create clusters for large subject pools, and help to better organize data in order to help create validity in a study. When looking at several different domains of self-efficacy and comparing these domains to different demographic information, it is useful to be able to cluster data so that variables can stay organized and the data is easier to track and less likely to become misconstrued. The purpose of the current study is to examine the relationship between several different domains of self-efficacy and participants' demographic information, as well as their overall well-being. Using k-means cluster analysis, the data for different domains of self-efficacy will be organized into clusters and then used to measure said domain's relationship to both itself and other variable data from the conducted study. By separating the data for each domain into cluster, it is hypothesized that the clusters will show to be significantly different from one other within the same domain, and the clusters will have participants of different backgrounds clustered together as well. For example, it is hypothesized that participants of a certain age range will be in one cluster, while another age range is in another cluster, and so on.

Methods

Participants

Participants for this study consisted of 92 Manhattan College undergraduate students, with ages ranging from 18 to 25 (M = 20.36, SD = 1.447). Of the 92 students polled, 75 (81.5%) were female and 17 (18.5%) were male. The data was collected from these students online using Qualtrics, and the students received no monetary value for participation.

Measures

In order to measure participants' academic self-efficacy, the Patterns of Adaptive Learning Survey (PALS: Midgley et al, 2000) was used. This is a five question, 5 point likert-type scale survey asked questions that specifically related to how confident a participant is in their ability to perform tasks in an academic setting, and has been validated to determine self-efficacy in the domain of academia with an alpha level of .78 (PALS: Midgley et al, 2000). To measure work self-efficacy and familial self-efficacy (the degree to which one believes they fulfill their role as a family member) in participants, an adaptation of Schwarzer and Jerusalem's General Self-Efficacy Scale (1995) was used. This scale contains ten questions in relation to one's general self-efficacy, which is a measurement of overall confidence in one's ability to perform daily tasks. All ten questions have 4-point likert-type scale responses. These scales were validated in order to measure general self-efficacy, with alpha levels ranging from .76 to .90, with a majority of scores being in the high .80s (Schwarzer & Jerusalem, 1995); however, each question was modified in order to be more specific to the domain in which data was being collected (work and familial self-efficacy, specifically). Finally, participants completed the Satisfaction with Life Scale (Diener, Emmons, Larsen & Griffin, 1985) in order to measure overall well-being in participants. This scale has five questions, with 7-point likert-type scale responses as well, and is validated to measure one's overall well-being, with a two-month test retest reliability coefficient of .82 and a coefficient alpha of .87 (Diener, Emmons, Larsen & Griffin, 1985).

Results

The data from the participants was split into clusters for each domain of self-efficacy measured in this study (self-efficacy at work, self-efficacy at home/with family, and self-efficacy in academics). Using the NbClust command in R, the optimal number of clusters was found for each of the three self-efficacy domains. The results of this analysis found both work specific self-efficacy and familial self-efficacy to have five clusters, while academic self-efficacy had two clusters. Using this information, a k-means clustering analysis was used in order to organize each of the three domains' results into the optimal number of clusters found from the previous NbClust command. This allowed the clusters to be created for each domain, so that the clusters could be used as individual, independent variables. The results of this k-means cluster grouped the data from work related self-efficacy into five clusters, the data from familial self-efficacy into five clusters, and the data from academic self-efficacy into only two clusters. For the work-related self-efficacy clusters, the first cluster consisted of 27 data points, the second cluster consisted of 20 data points, the third cluster consisted of 10 data points, the fourth cluster consisted of 21 data points, and the fifth cluster consisted of 14 data points. In the case of familial self-efficacy, the first cluster consisted of 25 data points, the second cluster consisted of 7 data points, the third cluster consisted of 33 data points, the fourth cluster consisted of 4 data point, and the fifth cluster consisted of 23 data points. Finally, the academic self-efficacy clusters had 52 data points in the first cluster and 40 data points in the second cluster.

Using the groups created by the cluster analysis, an ANOVA was run in order to examine the groups created by the clusters, in order to see if these created clusters are truly different from one another. The ANOVA for the five clusters created for the work-related self-efficacy data found a F value of 83.32 ($p < 2e^{-16}$). This shows that the clusters created by the k-means clustering method are significantly different from one another, implying that the clustering was successful and the groups represent significantly different data points across the likert-type scale. The ANOVA for the five clusters created for the familial self-efficacy data found a F value of 190.4 ($p < 2e^{-16}$). This result also shows how the created clusters are significantly different from one another, similar to the results of the work-related self-efficacy. Finally, the ANOVA for the two clusters created for the academic self-efficacy data found a F value of 218.9 ($p<2e^{-16}$). Just like the results of the previous cluster analyses, this implies the two groups created are significantly different from one another, and thus using NbClust to find the optimal number of clusters for each domain of self-efficacy, the k-means cluster analysis was successful in separating the data into relevant groups together.

Once the ANOVA was complete and it was established that the created clusters had validity, a MANOVA was run in order to compare several demographic factors to the clusters of the different domains of self-efficacy. First, a MANOVA was run using the three domains of self-efficacy measured in this study as a factor of the ethnicity of the participants in this study. The results for this MANOVA showed a F value of 1.0334 ($p=0.4183$). These results are not significant ($p>0.05$), meaning that participants' ethnicity did not have a relationship with or effect the group or cluster their self-efficacy data was put into. In a similar process, a MANOVA was run using participants' identified gender as a factor. The results of this MANOVA found a F value of 0.60924 ($p=0.6108$). This result would suggest that a participant's gender does not have a significant effect on which cluster their self-efficacy data is categorized into ($p>0.05$). For the last demographic test, a MANOVA was used with participants' age as a factor. The results of this MANOVA found a F value of 0.90926 (p=0.5794). These findings would indicate that a participant's age also does not have a significant impact on the cluster in which their self-efficacy data was grouped (p>0.05).

The last analysis examined in this study was a comparison of the clusters of the three domains of self-efficacy and participants' average overall well-being score. Once again, using a MANOVA, average overall well-being was used as a factor. The results of this MANOVA found a F value of 1.3079 (p=0.06946). This result indicates a moderate relationship between participants' well-being and which cluster their self-efficacy data was placed into (0.1>p>0.05).

Discussion

The purpose of this study was to utilize cluster analysis in order to examine relevant data. It was hypothesized that one's demographic information and one's overall well-being would be related to one's level of self-efficacy in three different domains, being self-efficacy at work, self-efficacy at home with family, and self-efficacy in academics. By utilizing NbClust in the program R, the optimal number of clusters for each domain of self-efficacy was determined. Using the information gathered from this, a k-means clustering analysis was conducted, and grouped data points from each domain into the optimal number of clusters given from the NbClust command. Finally, the clustered data was able to be used and treated as a group, so that the original hypothesis could be tested.

The results showed that the clustering process worked, where each cluster within its own domain was significantly different from one another. This shows how the process worked, considering the point of clustering is to separate the data given into different groups so that it can be examined and compared in study. This was done effectively for each domain, as was shown from the ANOVA in each domain. However, the hypothesis that demographic information would be related to groups of clusters in each domain of self-efficacy was not supported from this study. The MANOVA results reported that each of the three demographic questions used in this study had no significant relationship to any clusters in the domains of self-efficacy. Although it was not originally hypothesized, participants' overall well-being was also compared to clusters of self-efficacy though MANOVA. The results of this process did show a moderate relationship between the two. The p value was greater than 0.05, which is the generally accepted standard for significance, however the p value was less than 0.1, which can be used as the standard for significance in some cases. Due to the p value falling in between these two levels of significance, it can be argued that well-being does have some relation to self-efficacy, as previous research has noted.

Future research should focus on this relationship between self-efficacy in different domains of life and overall well-being more in depth. In addition to the examination of this concept, future research could work to include different types of clustering in order to see the results that come from different clusters of data. This could be an interesting concept for study that could lead researchers to different conclusions and interpretations of the dataset provided in this study.

Although the overall hypothesis was not supported, this study exemplified the use of clustering and cluster analysis, where data was clustering into groups and shown to give a significantly different selection of independent groups. Clustering data into these groups opens the door for more exploratory research, and allows researchers to examine the data with greater depth and understanding of the underlying implications given from the dataset. Being able to understand the benefits of cluster analysis is an extremely helpful tool and allows scientists and researchers alike to be able to extract more meaning from data and datasets, and to better understand the information given from a poll or survey.

Reference

Cornish, R. (2007). Statistics: Cluster Analysis. *Mathematics Learning Support Centre.* Retrieved from http://www.statstutor.ac.uk/resources/uploaded/clusteranalysis.pdf

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. Journal of Personality Assessment, 49, 71-75.

Gist, M. E. (1987). Self-Efficacy: Implications for Organizational Behavior and Human Resource Management. The Academy of Management Review, 12(3). pp. 472-485.

Kassambara, A. (2018). Clustering Distance Measures. Retrieved from https://www.datanovia.com/en/lessons/clustering-distance-measures/#methods-for-measuring-distances

Lent, R. W., Brown, S. D., & Gore Jr., P. A. (1997). Discriminant and Predictive Validity of Academic Self-Concept, Academic Self-Efficacy, and Mathematics-Specific Self-Efficacy. Journal of Counseling Psychology, 44(3). pp. 307-315

Loprinzi, P. D., & Walker, J. F. (2016). Association of Longitudinal Changes of Physical Activity on Smoking Cessation Among Young Daily Smokers. *Journal of Physical Activity and Health, 13*. Retrieved from http://web.a.ebscohost.com.proxy.bsu.edu/ehost/pdfviewer/pdfviewer?vid=11&sid=98ae2aa5-39af-4642-bff9-8b071ab29b86%40sessionmgr4009

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). Manual for the Patterns of Adaptive Learning Scales. Ann Arbor, MI: University of Michigan.

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. Measures in health psychology: A user's portfolio. Causal and control beliefs (pp. 35-37)

Suganya, R. & Shanthi, R. (November, 2012). Fuzzy C-Means Algorithm – A Review. *International Journal of Scientific and Research Publications, 2*(11). Retrieved from http://www.ijsrp.org/research-paper-1112/ijsrp-p1168.pdf

Tong, Y., Song, S. (2004). A Study on General Self-Efficacy and Subjective Well-Being of Low SES College Students in a Chinese University. College Student Journal, 28(4). pp. 637-642