2015 ■ vol. 11 ■ no. 1

DOI: 10.20982/tqmp.11.1.p008

Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data

Odilia Yim [a], Kylee T. Ramdeen [a, b, c]

[a] School of Psychology, University of Ottawa

[b] Laboratoire de Psychologie et Neurocognition, Université de Savoie

[c] Centre National de la Recherche Scientifique, Laboratoire de Psychologie et Neurocognition, Unité Mixte de Recherche 5105, Grenoble, France

Abstract ■ Cluster analysis refers to a class of data reduction methods used for sorting cases, observations, or variables of a given dataset into homogeneous groups that differ from each other. The present paper focuses on hierarchical agglomerative cluster analysis, a statistical technique where groups are sequentially created by systematically merging similar clusters together, as dictated by the distance and linkage measures chosen by the researcher. Specific distance and linkage measures are reviewed, including a discussion of how these choices can influence the clustering process by comparing three common linkage measures (single linkage, complete linkage, average linkage). The tutorial guides researchers in performing a hierarchical cluster analysis using the SPSS statistical software. Through an example, we demonstrate how cluster analysis can be used to detect meaningful subgroups in a sample of bilinguals by examining various language variables.

Keywords ■ Cluster analysis; Hierarchical cluster analysis; Agglomerative, linkage; SPSS

odilia.yim@uottawa.ca

Introduction

In everyday life, we try to sort similar items together and classify them into different groups, a natural and fundamental way of creating order among chaos. Among many scientific disciplines, it is also essential to uncover similarities within data to construct meaningful groups. The purpose of cluster analysis is to discover a system of organizing observations where members of the group share specific properties in common. Cluster analysis is a class of techniques that classifies cases into groups that are relatively homogeneous within themselves and relatively heterogeneous between each other (Landau & Chis Ster, 2010; Norusis, 2010). Cluster analysis has a simple goal of grouping cases into homogeneous clusters, yet the choice in algorithms and measures that dictates the successive merging of similar cases into different clusters makes it a complex process. Although an appealing technique, cluster solutions can be easily misinterpreted if the researcher does not fully understand the procedures of cluster analysis. Most importantly, one must keep in mind that cases will always be grouped into clusters regardless of the true nature of the data. Therefore, the present paper aims to provide researchers a background to hierarchical cluster analysis and a tutorial in SPSS using an example from psychology.

Cluster analysis is a type of data reduction technique. Data reduction analyses, which also include factor analysis and discriminant analysis, essentially reduce data. They do not analyze group differences based on independent and dependent variables. For example, factor analysis reduces the number of factors or variables within a model and discriminant analysis classifies new cases into groups that have been previously identified based on specific criteria. Cluster analysis is unique among these techniques because its goal is to reduce the number of cases or observations[1] by classifying them into homogeneous clusters, identifying groups without previously knowing group membership or the number of possible groups. Cluster analysis also allows for many options regarding the algorithm for combining groups, with each choice

[1] The present paper focuses only on the grouping of cases or observations, but cluster analysis can also be used to reduce the number of variables in a dataset.

resulting in a different grouping structure. Therefore, cluster analysis can be a convenient statistical tool for exploring underlying structures in various kinds of datasets.

Cluster analysis was initially used within the disciplines of biology and ecology (Sokal & Sneath, 1963). Although this technique has been employed in the social sciences, it has not gained the same widespread popularity as in the natural sciences. A general interest in cluster analysis increased in the 1960s, resulting in the development of several new algorithms that expanded possibilities of analysis. It was during this period that researchers began utilizing various innovative tools in their statistical analyses to uncover underlying structures in datasets. Within a decade, the growth of cluster analysis and its algorithms reached a high point. By the 1970s, the focus shifted to integrating multiple algorithms to form a cohesive clustering protocol (Wilmink & Uytterschaut, 1984). In recent decades, there has been a gradual incorporation of cluster analysis into other areas, such as the health and social sciences. However, the use of cluster analysis within the field of psychology continues to be infrequent (Borgen & Barnett, 1987).

The general technique of cluster analysis will first be described to provide a framework for understanding hierarchical cluster analysis, a specific type of clustering. The multiple parameters that must be specified prior to performing hierarchical clustering will be examined in detail. A particular focus will be placed on the relative impact of three common linkage measures. The second part of this paper will illustrate how to perform a hierarchical cluster analysis in SPSS by applying the technique to differentiate subgroups within a group of bilinguals. This paper will discuss the statistical implications of hierarchical clustering and how to select the appropriate parameters in SPSS to allow researchers to uncover the grouping structure that most accurately describes their multivariate dataset.

Hierarchical Cluster Analysis

Due to the scarcity of psychological research employing the general technique of cluster analysis, researchers may not fully understand the utility of cluster analysis and the application of the clustering technique to their data. There are two main methods: hierarchical and non-hierarchical cluster analysis. Hierarchical clustering combines cases into homogeneous clusters by merging them together one at a time in a series of sequential steps (Blei & Lafferty, 2009). Non-hierarchical techniques (e.g., *k*-means clustering) first establish an initial set of cluster means and then assign each case to the closest cluster mean (Morissette & Chartier, 2013). The present paper focuses on hierarchical clustering, though both clustering methods have the same goal of increasing within-group homogeneity and between-groups heterogeneity. At each step in the hierarchical procedure, either a new cluster is formed or one case joins a previously grouped cluster. Each step is irreversible meaning that cases cannot be subsequently reassigned to a different cluster. This makes the initial clustering steps highly influential because the first clusters generated will be compared to all of the remaining cases. The alternate method of non-hierarchical clustering requires the researcher to establish a priori the number of clusters in the final solution. If there is uncertainty about the total number of clusters in the dataset, the analysis must be re-run for each possible solution. In this situation, hierarchical clustering is preferred as it inherently allows one to compare the clustering result with an increasing number of clusters; no decision about the final number of clusters needs to be made a priori.

Hierarchical cluster analysis can be conceptualized as being agglomerative or divisive. Agglomerative hierarchical clustering separates each case into its own individual cluster in the first step so that the initial number of clusters equals the total number of cases (Norusis, 2010). At successive steps, similar cases–or clusters–are merged together (as described above) until every case is grouped into one single cluster. Divisive hierarchical clustering works in the reverse manner with every case starting in one large cluster and gradually being separated into groups of clusters until each case is in an individual cluster. This latter technique, divisive clustering, is rarely utilized because of its heavy computational load (for a discussion on divisive methods, see Wilmink & Uytterschaut, 1984). The focus of the present paper is on the method of hierarchical agglomerative cluster analysis and this method is defined by two choices: the measurement of distance between cases and the type of linkage between clusters (Bratchell, 1989).

Distance Measure

The definition of cluster analysis states it is a technique used for the identification of homogeneous subgroups. Therefore, cluster analysis is inherently linked to the

concept of similarity. The first step a researcher must take is to determine the statistic that will be used to calculate the distance or similarity between cases. Both measures may be thought to mirror one another; as the distance between two cases decreases, their similarity should respectively increase. However, an important distinction must be made: whereas both measures reflect the pattern of scores of the chosen variables, only the distance measure takes into account the elevation of those scores (Clatworthy, Buick, Hankins, Weinman, & Horne, 2005). For example, if we wish to separate bilinguals who switch between their two languages frequently from those who do not switch languages often, the difference in the actual scores on multiple language measures must be taken into account. In this case, a distance measure must be selected. However, if we wish to assess the efficacy of a language intervention program, then the actual language scores may not be of importance. In this case, we would be assessing the pattern of language scores over time (i.e. from before to after intervention) to identify the clusters of people that improved, worsened, or did not change their language skills after intervention. In this situation, a similarity measure such as the Pearson correlation, would be sufficient to assess the pattern of scores before and after intervention while ignoring the raw language scores. An added difficulty of using a correlation coefficient is that it is easy to interpret when there are only one or two variables, but as the number of variables increases the interpretation becomes unclear. It is for these reasons that distance measures are more commonly used in cluster analysis because they allow for an assessment of both the pattern and elevation of the scores in question.

Of course, there is not only one statistic that can be used as a distance measure in cluster analysis. The choice of the distance measure will depend primarily on whether the variables are continuous or dichotomous in nature. Many chapters on cluster analysis simply overlook this question and discuss measures applicable to continuous variables only. Although this paper will focus on applying cluster analysis to continuous data, it is important to note that at least four measures exist for calculating distance with dichotomous data (see Finch, 2005).

The most commonly used distance measure for continuous variables is the squared Euclidean distance, $\sum_{j=1}^{k}(a_j - b_j)^2$. In the equation, *a* and *b* refer to the two cases being compared on the *j* variable, where *k* is the total number of variables included in the analysis (Blei & Lafferty, 2009). This algorithm allows for the distance between two cases to be calculated across all variables and reflected in a single distance value. At each step in the procedure, the squared Euclidean distance between all pairs of cases and clusters is calculated and shown in a proximity matrix (discussed below). At each step, the pair of cases or clusters with the smallest squared Euclidean distance will be joined with one another. This makes hierarchical clustering a lengthy process because after each step, the full proximity matrix must once again be recalculated to take into account the recently joined cluster. The squared Euclidean distance calculation is straightforward when there is only one case per cluster. However, an additional decision must be made as to how best to calculate the squared Euclidean distance when there is more than one case per cluster. This is referred to as the linkage measure and the researcher must determine how to best calculate the link between two clusters.

Linkage Measure

The problem that arises when a cluster contains more than one case is that the squared Euclidean distance can only be calculated between a pair of scores at a time and cannot take into account three or more scores simultaneously. In line with the proximity matrix, the goal is still to calculate the difference in scores between pairs of clusters, however in this case the clusters do not contain one single value per variable. This suggests that one must find the best way to calculate an accurate distance measure between pairs of clusters for each variable when one or both of the clusters contains more than one case. Once again, the goal is to find the two clusters that are nearest to each other in order to merge them together. There exist many different linkage measures that define the distance between pairs of clusters in their own way. Some measures define the distance between two clusters based on the smallest or largest distance that can be found between pairs of cases (single and complete linkage, respectively) in which each case is from a different cluster (Mazzocchi, 2008). Average linkage averages all distance values between pairs of cases from different clusters. Single linkage, complete linkage, and average linkage will each be fully detailed in turn.

Single linkage. Also referred to as nearest neighbour or minimum method. This measure defines the distance between two clusters as the minimum distance found between one case from the first cluster and one case

from the second cluster (Florek, Lukaszewiez, Perkal, Steinhaus, & Zubrzchi, 1951; Sneath, 1957). For example, if cluster 1 contains cases *a* and *b*, and cluster 2 contains cases *c*, *d*, and *e*, then the distance between cluster 1 and cluster 2 would be the smallest distance found between the following pairs of cases: (*a*, *c*), (*a*, *d*), (*a*, *e*), (*b*, *c*), (*b*, *d*), and (*b*, *e*). A concern of using single linkage is that it can sometimes produce chaining amongst the clusters. This means that several clusters may be joined together simply because one of their cases is within close proximity of case from a separate cluster. This problem is specific to single linkage due to the fact that the smallest distance between pairs is the only value taken into consideration. Because the steps in agglomerative hierarchical clustering are irreversible, this chaining effect can have disastrous effects on the cluster solution.

Complete linkage. Also referred to as furthest neighbour or maximum method. This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest distance between pairs of cases (Sokal & Michener, 1958). Although this solves the problem of chaining, it creates another problem. Imagine that in the above example cases *a*, *b*, *c*, and *d* are within close proximity to one another based upon the pre-established set of variables; however, if case *e* differs considerably from the rest, then cluster 1 and cluster 2 may no longer be joined together because of the difference in scores between (*a*, *e*) and (*b*, *e*). In complete linkage, outlying cases prevent close clusters to merge together because the measure of the furthest neighbour exacerbates the effects of outlying data.

Average linkage. Also referred to as the Unweighted Pair-Group Method using Arithmetic averages (UPGMA)[2]. To overcome the limitations of single and complete linkage, Sokal and Michener (1958) proposed taking an average of the distance values between pairs of cases. This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters. For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged. This means that in the previous example, the distance between cluster 1 and cluster 2 would be the average of all distances between the pairs of cases listed above: (*a*, *c*), (*a*, *d*), (*a*, *e*), (*b*, *c*), (*b*, *d*), and (*b*, *e*). Incorporating information about the variance of the distances renders the average distance value a more accurate reflection of the distance between two clusters of cases.

Each linkage measure defines the distance between two clusters in a unique way. The selected linkage measure will have a direct impact on the clustering procedure and the way in which clusters are merged together (Mazzocchi, 2008). This will subsequently impact the final cluster solution. In the next section, a hierarchical cluster analysis will be performed on a previously published dataset using SPSS.

SPSS Tutorial on Hierarchical Cluster Analysis

The following tutorial will outline a step-by-step process to perform a hierarchical cluster analysis using SPSS statistical software (version 21.0) and how to interpret the subsequent analysis results. The research data in the following example was part of a larger research dataset from Yim and Bialystok (2012) which examined bilinguals and their language use. The present example includes 67 Cantonese-English bilingual young adults. The participants completed language proficiency tests in both languages and questionnaires regarding their daily language use. Participants had to indicate how often they use both English and Cantonese daily ("I use English and Cantonese daily") on a scale from 0 (*none of the time*) to 100 (*all of the time*). Language proficiency was assessed using the Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997) in both Cantonese and English, measuring receptive vocabulary. This sample was chosen because it is an apt example to demonstrate the applicability of cluster analysis on psychological data. Bilinguals are loosely defined as individuals who regularly use two (or more) languages, yet many issues remain in the research field; for example, there is still no consensus as to what criteria determine that someone is bilingual (Grosjean, 1998). High proficiency bilinguals are often viewed as a homogenous population; however, there can be within-group differences in language usage and language proficiency. The goal of a hierarchical cluster analysis on this data is to examine possible subgroups in a sample of highly proficient bilinguals.[3]

[2] The average linkage presented here is referred to as average linkage between groups in SPSS and other resources. It should not be confused with an alternate method, average linkage within groups, which takes into account the variability found within each cluster. For a contrast between linkage measures, see Everitt, Landau, Leese, and Stahl (2011).

[3] To practice with a dataset, please contact the corresponding

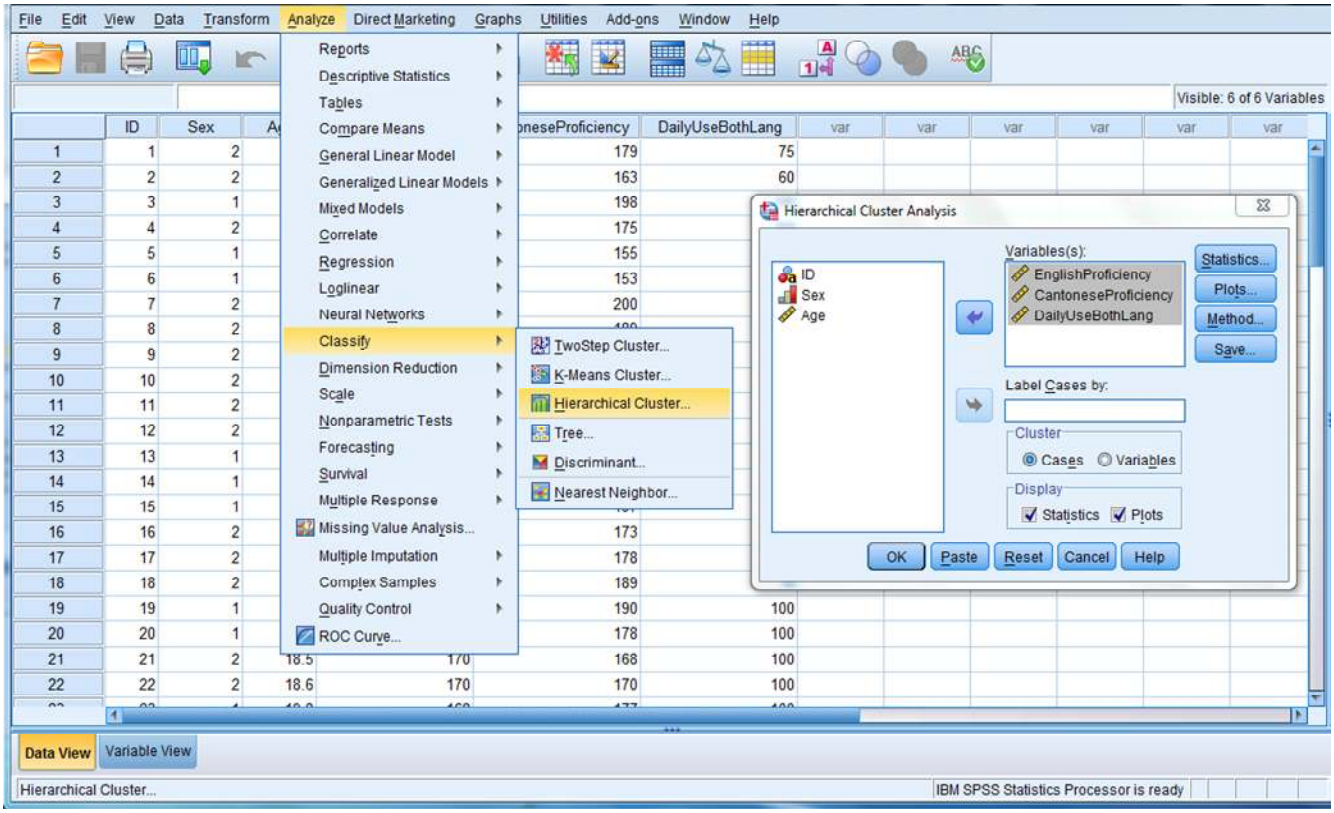

Figure 1 ■ Running a hierarchical cluster analysis.

Step 1: Choosing Cluster Variables

The researcher first has to identify the variables that will be included for analysis. Any number of variables can be included, but it is best to include variables that are meaningful to the research question. In this example, we use three variables for the cluster analysis: bilinguals' proficiency scores in both languages and their self-report of their daily use of both languages. These three variables target proficiency and daily use, two dimensions commonly used to assess bilingualism. The variables included in this example are all continuous variables.

Step 2: Selecting Cluster Method

To run a hierarchical cluster analysis in SPSS, click on Analyze, then Classify, and then Hierarchical Cluster (Figure 1). A new dialog box labelled Hierarchical Cluster Analysis will then appear. Among the list of variables presented in the left panel, select the variables that will be included in the analysis and move them to the Variables box on the right. As shown in Figure 1, the three selected language variables have been moved into the Variables box. There is also an option to label cases. If a researcher has a variable which can be used to identify the individual cases, the variable can be brought over to the box named Label Cases By. This can be helpful in reading the output as it will allow for each case to be easily referenced. In our example, we do not assign a variable to label cases because the participant ID numbers correspond with the row numbers in SPSS. If no variable is chosen to label the cases, the output will use the SPSS row numbers to identify the cases.

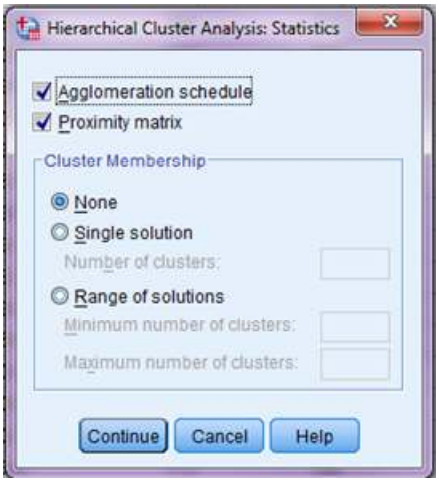

Figure 2 ■ Choosing statistics.

author.

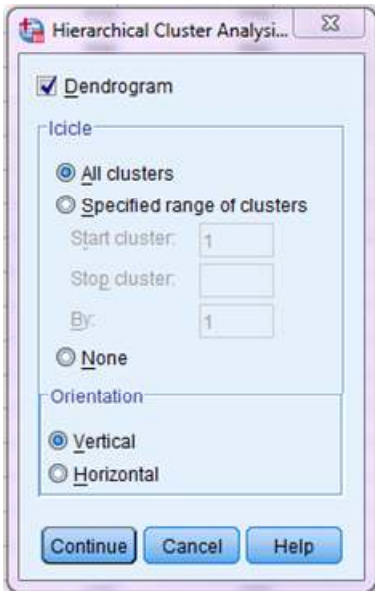

Figure 3 ■ Selecting plot options.

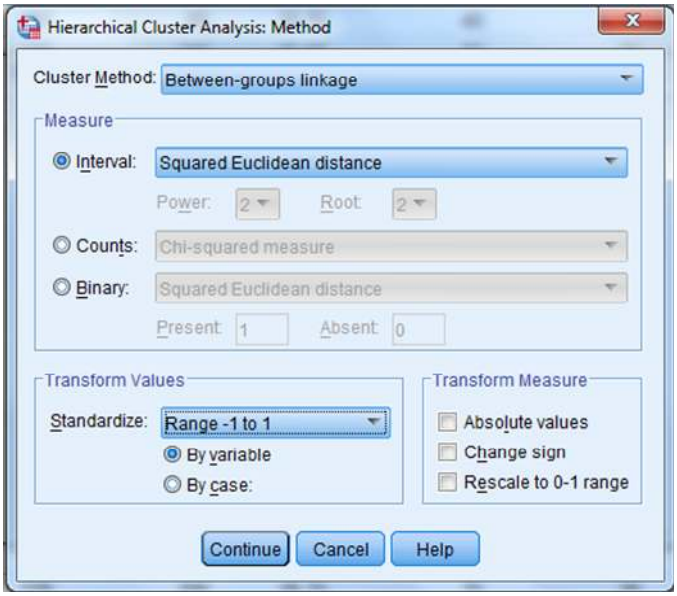

Figure 4 ■ Specifying cluster measures.

Step 3: Specifying Parameters

After selecting the variables to include in the analysis, it is important to request certain items to be included in the data output. On the right side of the window, the Statistics button will allow for the researcher to request a proximity matrix to be produced (Figure 2). Otherwise, the SPSS output will only include the agglomeration schedule, a table that details each step of the clustering procedure.

Under Statistics, the Plots button allows the researcher to select the visual outputs that will illustrate the cluster solution. The researcher can choose to produce a dendrogram, a visual tree graph that displays the clustering procedure. Dendrograms are very helpful in determining where the hierarchical clustering procedure should be stopped because the ultimate goal is not to continue the clustering until each case is combined into one large cluster. In Plots, the box marked Dendrograms must be selected otherwise SPSS will not generate it automatically (Figure 3). Also, the researcher can modify the presentation of the icicle plot by changing its orientation.

Next, it is important to set the specific parameters for the cluster analysis, namely choosing the distance and linkage measures that will be used. By clicking on the Method button, a new dialog box will open where these options will be listed (Figure 4). The Cluster Method refers to the linkage measure. In SPSS, the default is the between-groups linkage which is equivalent to average linkage between groups. In the drop-down menu, single linkage and complete linkage are also available along with four other measures. Under the Measure options, Interval specifies the distance measure for the cluster analysis. The default option is the squared Euclidean distance as it is the most common and some linkage measures specifically require this distance measure. SPSS provides eight distance options, including Euclidean distance and Pearson correlation.

Under Transform Values, there are options to standardize the variables selected for clustering. No standardization is specified by default, but the two most common transformation options are z-scores or using a range of -1 to 1. In our example, the values will need to be transformed as the three variables were not measured on the same scale. Now that all the parameters have been set and the output options have been chosen, the analysis is ready to be run in SPSS. The SPSS syntax for the tutorial can also be found in the Appendix.

Step 4: Interpreting the Output

Similar to other analyses, SPSS will first produce a Case Processing Summary which lists the number of valid cases, the number of missing cases, and also the distance measure that was chosen (i.e., the squared Euclidean distance). The Proximity Matrix is the second table in the output, if requested. The matrix lists the squared Euclidean distance that was calculated

Table 1 ■ Proximity Matrix

Case	Squared Euclidean Distance										
	Case 50	Case 51	Case 52	Case 53	Case 54	Case 55	Case 56	Case 57	Case 58	Case 59	Case 60
Case 50	.000	.499	.145	.504	1.404	.162	.162	.114	.222	.933	.132
Case 51	.499	.000	.362	.390	.360	.278	.278	.256	.252	.116	.270
Case 52	.145	.362	.000	.740	1.360	.028	.028	.282	.115	.867	.058
Case 53	.504	.390	.740	.000	.842	.753	.753	.329	.803	.379	.703
Case 54	1.404	.360	1.360	.842	.000	1.104	1.104	.760	.900	.140	1.022
Case 55	.162	.278	.028	.753	1.104	.000	.000	.208	.029	.737	.009
Case 56	.162	.278	.028	.753	1.104	.000	.000	.208	.029	.737	.009
Case 57	.114	.256	.282	.329	.760	.208	.208	.000	.176	.492	.144
Case 58	.222	.252	.115	.803	.900	.029	.029	.176	.000	.660	.015
Case 59	.933	.116	.867	.379	.140	.737	.737	.492	.660	.000	.700
Case 60	.132	.270	.058	.703	1.022	.009	.009	.144	.015	.700	.000

between all pairs of cases in the first step of the cluster procedure. Table 1 is a truncated version of the matrix that shows the distances between cases 50-60; in the example, cases 55 and 56 had the smallest squared Euclidean distance (approximately .000) and were therefore the first two cases to be joined together. The full proximity matrix is recalculated after each step but is not shown in the output to save space. Nonetheless, the repeated calculation of the proximity matrix is used to determine the successive merging of cases illustrated in the remaining outputs.

The Agglomeration Schedule (Table 2) follows the proximity matrix in the output. The agglomeration schedule displays how the hierarchical cluster analysis progressively clusters the cases or observations. Each row in the schedule shows a stage at which two cases are combined to form a cluster, using an algorithm dictated by the distance and linkage selections. The agglomeration schedule lists all of the stages in which the clusters are combined until there is only one cluster remaining after the last stage. The number of stages in the agglomeration schedule is one less than the number of cases in the data being clustered. In this example, there are 66 stages because the sample consists of 67 bilinguals. The coefficients at each stage represent the distance of the two clusters being combined. As shown in Table 2, cases 55 and 56 are combined at the first stage because the squared Euclidean distance between them is the smallest out of all the pairs. In fact, the coefficients are very small (approximately .000) for the first several stages and slowly increase as the schedule progresses. The increase in coefficients indicates that the clusters being combined at a given stage are more heterogeneous than previous combinations. (The agglomeration schedule shown in Table 2 has been cropped. Only the top and the bottom of the schedule are shown as it becomes quite long with a large number of cases.)

The purpose of the agglomeration schedule is to assist the researcher in identifying at what point two clusters being combined are considered too different to form a homogeneous group, as evidenced by the first large increase in coefficient values. When there is a large difference between the coefficients of two consecutive stages, this suggests that the clusters being merged are increasing in heterogeneity and that it would be ideal to stop the clustering process before the clusters become too dissimilar. In Table 2, there is a jump in the coefficient values between stages 63 and 64. With a difference of approximately .201, this is the first noticeable increase that we encounter as we move down the list of coefficients in the agglomeration schedule. Therefore, we can choose to stop the clustering after stage 63.

It can be difficult to calculate the differences of the coefficients. An easy solution is to plot the coefficient values by stage in a scree plot. A scree plot is simply a line graph, a visual representation of the agglomeration schedule. Although SPSS does not produce the scree plot in its output, it can be made in Microsoft Excel by copying the values in the stage and coefficients columns. In Figure 5, the scree plot shows a large increase in the coefficients after stage 63.

Table 2 ■ Agglomeration Schedule.

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	55	56	.000	0	0	25
2	9	32	.000	0	0	11
3	12	43	.000	0	0	6
4	3	46	.000	0	0	57
5	25	31	.000	0	0	8
6	12	19	.000	3	0	11
7	20	50	.001	0	0	17
8	25	28	.001	5	0	31
9	10	62	.001	0	0	18
10	21	24	.001	0	0	13
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
55	6	40	.082	41	45	59
56	4	13	.089	49	35	61
57	3	8	.092	4	52	63
58	1	2	.109	51	27	65
59	5	6	.116	34	55	61
60	15	51	.133	54	0	62
61	4	5	.167	56	59	64
62	15	18	.262	60	53	65
63	3	7	.303	57	50	64
64	3	4	.504	63	61	66
65	1	15	.603	58	62	66
66	1	3	.887	65	64	0

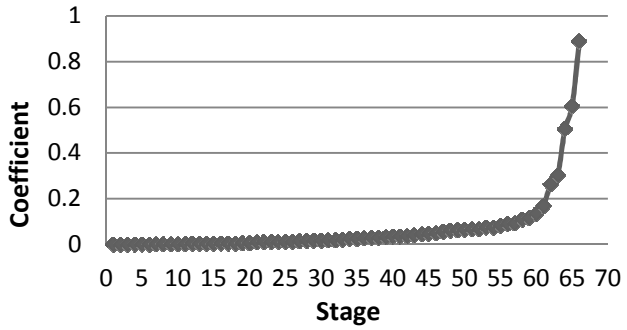

Figure 5 ■ Scree plot of coefficients by stage.

The first figure included in the SPSS output is the Icicle Plot (Figure 6). Like the agglomeration schedule, this plot displays the similarity between two cases. The icicle plot is easier to interpret when examining it from the bottom to the top. Each of the dark grey bars in the plot represents one case. However, it is important to note the areas between cases and when they become shaded. The point at which the space between two cases becomes shaded represents when the cases were joined together. For example in Figure 6, near the midpoint of the plot, the section between two dark bars is shaded immediately, suggesting that those two cases were clustered together at the onset of the clustering procedure. Inspecting the plot closely, we discover that those two cases correspond to case 55 and case 56, which were combined at the first stage of the agglomeration schedule. (In the SPSS output, the bars on the icicle plot are all shaded in the same colour. We have changed the bars representing the cases into a darker colour to differentiate them more easily.)

As mentioned previously, a hierarchical cluster analysis is best illustrated using a dendrogram, a visual display of the clustering process (Figure 7). It appears at the very end of the SPSS output. Examining the dendrogram from left to right, clusters that are more similar to each other are grouped together earlier. The vertical lines in the dendrogram represent the grouping of clusters or the stages of the agglomeration schedule. They also indicate the distance between two joining clusters (as represented by the x-axis, located above the plot). As the clusters being merged become more heterogeneous, the vertical lines will be located farther to the right side of the plot, as they represent larger distance values. While the vertical lines are indicative of the distance between clusters, the horizontal lines

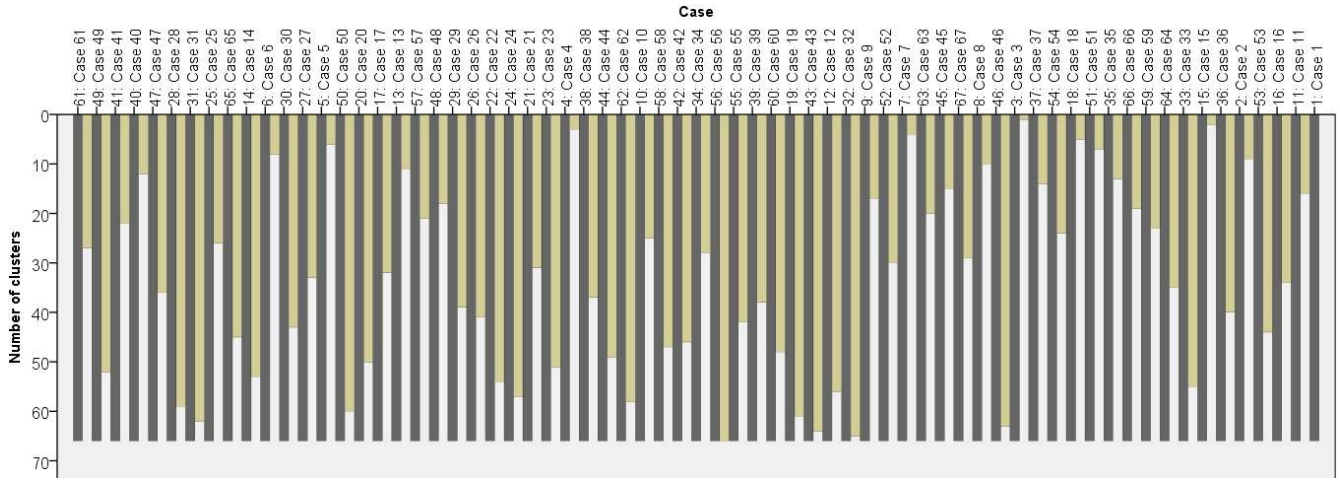

Figure 6 ■ Icicle plot.

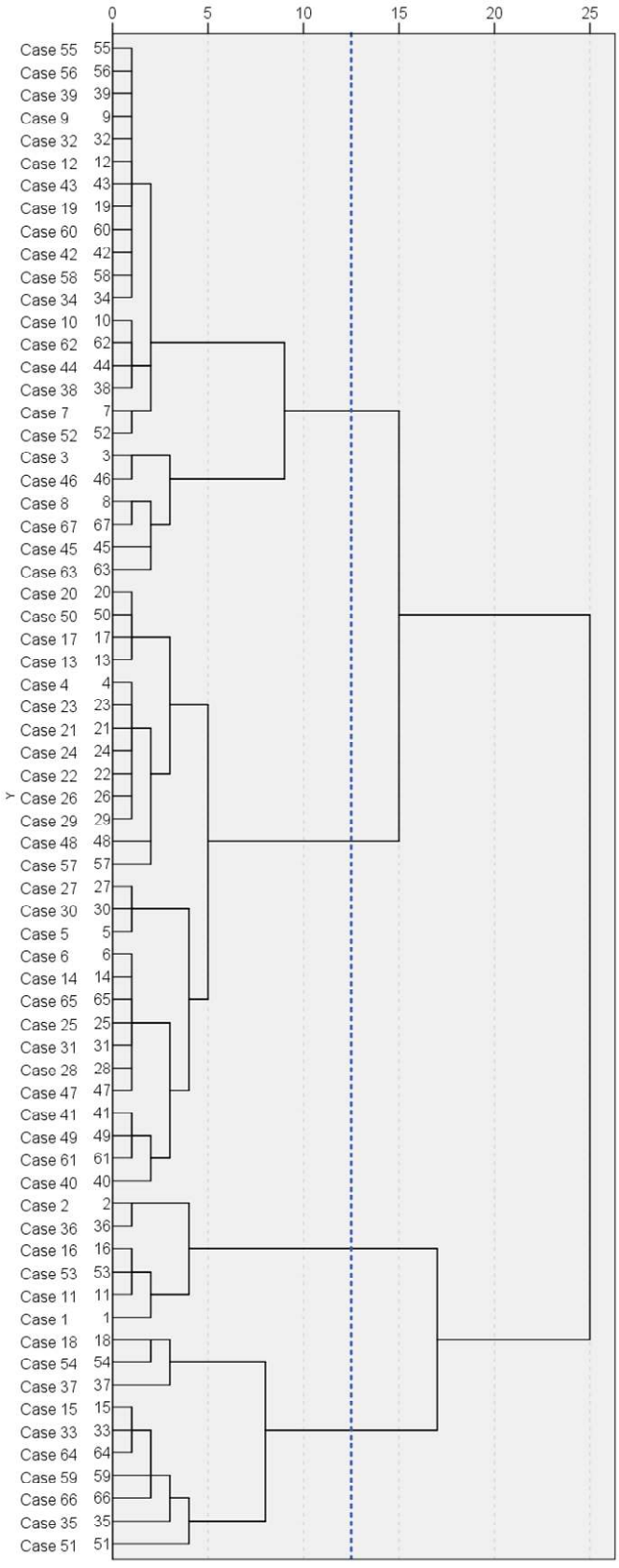

Figure 7 ■ Dendrogram with added line indicating suggested stopping location.

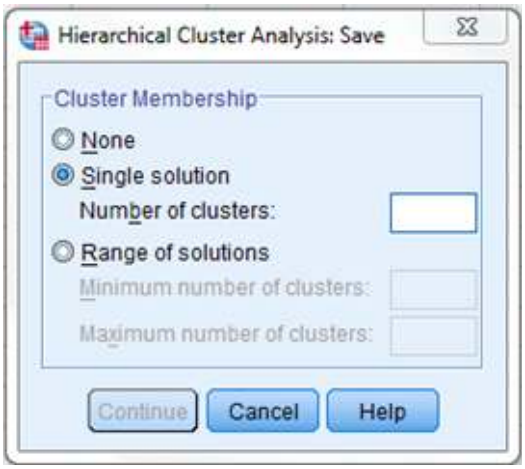

Figure 8 ■ Creating a cluster filter variable.

represent the differences of these distances. The horizontal lines also connect all cases that are a part of one cluster which is important when determining the final number of clusters after the stopping decision is made. Upon visually inspecting the dendrogram, the longest horizontal lines represent the largest differences. Therefore, a long horizontal line indicates that two clusters (which are dissimilar to each other) are being combined and identifies where it is optimal to stop the clustering procedure. Similar to the agglomeration schedule, if the vertical and horizontal lines are close to one another, then this would suggest that the level of homogeneity of the clusters merged at those stages is relatively stable. The cut-off should thus be placed where there are no closely plotted lines while eliminating the vertical lines with large values.

As there is no formal stopping rule for hierarchical cluster analysis, a cut-off needs to be determined from the dendrogram to signify when the clustering process should be stopped (Bratchell, 1989). The best approach to determine the number of clusters in the data is to incorporate information from both the agglomeration schedule and the dendrogram. Figure 7 illustrates the dendrogram generated by SPSS with an added line indicating the optimal stopping point of the clustering procedure. From the agglomeration schedule, we had concluded that it would be best to stop the cluster analysis after the 63rd stage, eliminating the last three stages (stages 64, 65, and 66). This decision is reflected in the dendrogram where the last three vertical lines (representing the last three stages in the agglomeration schedule) were cut from the cluster solution. By stopping the clustering at this point, four clusters are

revealed within the dataset as the cut-off line crosses four horizontal lines. The interpretation of these clusters will be discussed following the tutorial.

Step 5: Organizing Data into Subgroups

Once the number of clusters has been decided, the data can be organized into the subgroups specified by the analysis. This can be easily accomplished by re-running the hierarchical cluster analysis, but with one additional step. In the Hierarchical Cluster Analysis window, click on the button on the right called Save (under Method where we chose the cluster measures). As seen in Figure 8, the researcher can dictate the cluster membership to have a single solution (fixed number of clusters) or a range of solutions (a range of clusters). By default, SPSS does not specify cluster membership because it contradicts the objective of hierarchical clustering (i.e. not requiring a known number of clusters beforehand). Since we have determined the number of clusters in the data, we are able to request a specific number of clusters. In our example, four clusters were identified. By specifying the number of clusters in this Save window, SPSS will generate a new variable in the Data View window which assigns each case into one of the four clusters. This can also be accomplished by inserting a Save Cluster instruction in the SPSS syntax (see Appendix for syntax). Once the analysis is complete, the researcher is able to use the cluster variable to analyze the different clusters, for example, examining descriptive statistics and how the clusters may differ according to the variables used in the analysis.

Both windows allow the researcher to specify cluster membership, but it is only in the Save option where a cluster filter variable will be generated. Also, if the researcher is uncertain about the number of clusters in the data and wishes to look at two or more options, inputting a range of solutions can be used to generate a new variable for each of the cluster membership options.

Discussion

Cluster analysis allows the researcher to make many decisions about the measures used in the analysis. However, this can be a problem as it places greater weight on the researcher being knowledgeable to select the appropriate measures. The tutorial demonstrated that it is often difficult to determine the exact number of clusters in a dataset and that this decision is dependent on a numerical and visual inspection of the output figures which can sometimes be subjective and ambiguous. The underlying structure of the cluster solution can change greatly by simply modifying one of the chosen measures, such as linkage. The following section will review how employing three different linkage measures (single linkage, complete linkage, and average linkage) can result in three vastly different analyses and clustering results, as evidenced in visual plots such as dendrograms. Additionally, upon choosing a linkage measure, we interpret the results of the cluster solution and the meaning of the subgroups.

As mentioned previously, the linkage measure determines how to calculate the distance between pairs of clusters with two or more cases. Figure 9 displays three dendrograms from three analyses, each using a different linkage measure. Although all three analyses were run on the same data (from the SPSS tutorial), the differences between the dendrograms are easily observable upon visual inspection. First, the analysis using the single linkage measure is shown on the left. Using the process outlined in the tutorial, three clusters can be identified in the data. However, the dendrogram clearly shows how single linkage can produce chaining because the majority of cases were grouped together into a large cluster, with minimal distance between clusters. As the smallest distance between pairs is the only value taken into consideration, cases that are close in distance but from different clusters may drive their respective groups to merge despite the proximity of the rest of the cases. The dendrogram in the center shows the analysis using complete linkage where the opposite problem can be observed. Five clusters were derived from this analysis. Complete linkage does not necessarily merge groups that are close together due to outlying cases that may be far apart.

Average linkage represents a natural compromise between single linkage and complete linkage, as it is sensitive to the shape and size of clusters. Single linkage is sensitive to outliers, but it is impervious to differences in the density of the clusters; in contrast, complete linkage can break down large clusters though it is highly influenced by outliers (Almeida, Barbosa, Pais, & Formosinho, 2007). As seen by the visual comparison, the average linkage method was a compromise between the single and complete methods as well. (The dendrogram on the right in Figure 9 is the same as Figure 7.) However, the number of clusters obtained using average linkage is not always the average between the single and complete linkage solutions, as was the case in this example. Average

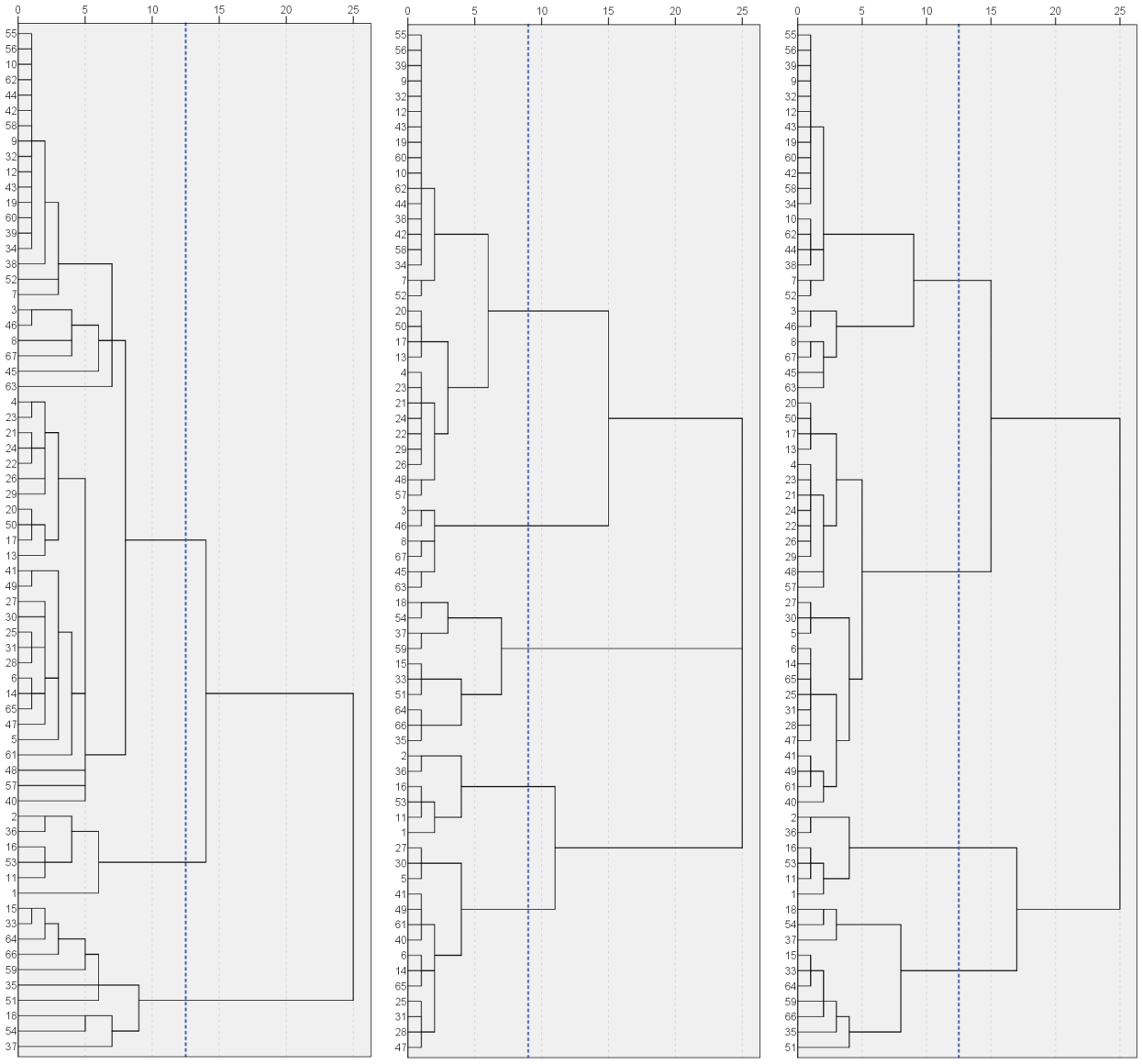

Figure 9 ■ Three dendrograms from a hierarchical cluster analysis with single linkage (left), complete linkage (center), and average linkage (right).

linkage was the most appropriate option for the data used in this example; however, the procedures and solution of a cluster analysis will be unique to each dataset. Bratchell (1989) suggests that there is no best choice and researchers may need to employ different techniques and compare their results.

It was demonstrated above that average linkage was the best linkage measure for the bilingual data in the present example. At this point, it is important to take a closer look at the cluster groups generated by the hierarchical cluster analysis and how these groups may be meaningful. The analysis resulted in creating four subgroups. As seen in Table 3, the analysis resulted in four distinct clusters that vary according to two dimensions, the daily use of both languages and Cantonese proficiency, as measured by the PPVT-III. Clusters A and B represent bilinguals who use Cantonese and English every day, while Clusters C and D are those who use both languages in a moderate degree only. When examining Cantonese proficiency, it is noteworthy that despite all bilinguals being communicatively competent in Cantonese, there is a split among them on this measure. The bilinguals in Clusters A and D obtained higher scores compared to

Table 3 ■ Means and standard deviations of daily language use and proficiency scores in Cantonese and English by cluster group.

Cluster	*n*	Daily Use of Both Languages	Cantonese Proficiency	English Proficiency
A	24	99.2 (2.5)	194.2 (4.3)	160.0 (19.2)
B	27	97.1 (5.0)	165.0 (9.0)	174.7 (10.2)
C	6	65.3 (5.5)	168.8 (7.5)	182.5 (4.2)
D	10	59.6 (7.5)	195.8 (4.5)	148.1 (17.5)

those in Clusters B and C. Therefore, four meaningful subgroups were detected: (i) Cluster A – frequent language users with high Cantonese proficiency; (ii) Cluster B – frequent language users with intermediate Cantonese proficiency, (iii) Cluster C – moderate language users with intermediate Cantonese proficiency, and (iv) Cluster D – moderate language users with high Cantonese proficiency. The results of the cluster analysis confirmed that there are meaningful subgroups within this group of high proficiency bilinguals. Although bilinguals are generally considered to be a homogeneous group, there exist fine differences among them and distinguishing these within-group differences can be significant for bilingual research.

After identifying a set of meaningful subgroups, there is still a final step that one can take to further validate the cluster solution by performing a split-sample validation.[4] The sample was randomly split to create two sub-samples, which were then used for comparison regarding the number of clusters and each of the cluster profiles (Everitt et al., 2011). The first sub-sample ($n = 34$) and the second sub-sample ($n = 33$) both generated the same four cluster groups as the original full-sample solution. The clustering pattern was maintained across the four subgroups: bilinguals who used Cantonese and English everyday (Clusters A and B) represented a larger proportion of both sub-samples than bilinguals who used their languages moderately (Clusters C and D). Importantly, the cluster solution was replicated within each of the four subgroups; that is, cases which were merged together in a cluster were also combined together in both of the sub-samples. Therefore, the split-sample validation technique complements our visual inspection of the cluster analysis and allows us to reliably conclude the existence of four meaningful subgroups within the dataset.

Deciding upon the most accurate cluster solution and its interpretation may in itself pose a limitation because of the freedom that is given to the researcher. Like with any other statistical analysis, there are situations in which hierarchical cluster analysis does not perform optimally. As explained above, the full proximity matrix must be computed at each step in the clustering procedure. If the sample is very large, more time will be needed to produce the proximity matrix at each step. Moreover, each step is irreversible and cases cannot be reassigned to a different cluster later on in the process. The sequential and inflexible nature of hierarchical clustering makes the initial partitions more influential than those at a later point. However, these potential limitations inherent to the nature of hierarchical cluster analysis are minimal and the benefits of this otherwise flexible method are broad and encouraging for its use as a statistical tool in the field of psychology.

Cluster analysis is not a data mining technique used for creating a structure within a dataset that is not meaningful. Hierarchical clustering will always provide a series of cluster solutions from one possible cluster to *n* possible clusters. The present paper does not consider comprehensively all the parameters associated with hierarchical cluster analysis; there are many specific techniques and models that have not been addressed. (We recommend the fifth edition of Cluster Analysis by Everitt et al., 2011, as further reading. It is a comprehensive and essential resource for researchers who are interested in this statistical technique.) It is the responsibility of the researcher to ensure that the distance and linkage measures have been appropriately selected and that the clustering process is stopped at the most logical point. As

[4] There are no cluster validation methods in SPSS; however, other validation techniques are available in different statistical software packages (SAS Institute, 1983; Wilkinson, Engelman, Corter, & Coward, 2000).

specified in the SPSS tutorial, the investigator must examine various outputs to determine the most appropriate number of clusters. There is no correct or incorrect solution to cluster analysis; it is up to the researcher to select the appropriate parameters to reveal the most accurate underlying structure of the data.

Conclusion

Cluster analysis is a statistical tool that offers a wide range of options for the researcher, allowing for the analysis to be uniquely tailored to the data and the objectives of the study. Although the practice of using this class of techniques is not yet common in the field of psychology, there are clear advantages to offering various options in setting the parameters of the analysis. Hierarchical cluster analysis is suggested as a practical method in identifying meaningful clusters within samples that may superficially appear homogeneous. The present paper presented a theoretical background to hierarchical clustering, specifically outlining the three common linkage measures used and a tutorial outlining the steps in the analysis, guiding researchers to discover underlying structures and subgroups on their own. With increased practice and when utilized appropriately, cluster analysis is a powerful tool that can be implemented on diverse sets of psychological data.

Authors' notes and acknowledgments

We would like to thank Sylvain Chartier for his suggestions on an earlier version of this paper and Ellen Bialystok for her guidance during the collection of the data used in the tutorial.

Address for Correspondence: Odilia Yim, University of Ottawa, 136 Jean Jacques Lussier, Vanier 5045, Ottawa, ON Canada K1N 6N5.

References

Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems, 87*, 208-217.

Blei, D. & Lafferty, J. (2009). Topic models. In A. Srivastava and M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71-94). Boca Raton, FL: Taylor & Francis Group.

Borgen, F. H. & Barnett, D. C. (1987). Applying cluster analysis in counselling psychology research. *Journal of Counseling Psychology, 34*(4), 456-468.

Bratchell, N. (1989). Cluster analysis. *Chemometrics and Intelligent Laboratory Systems, 6*, 105–125.

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.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test – III*. Circle Pines, MN: American Guidance Service.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis (5th edition)*. Chichester, UK: John Wiley & Sons, Ltd.

Finch, H. (2005). Comparison of distance measures in cluster analysis with dichotomous data. *Journal of Data Science*, *3*(1), 85-100.

Florek, K., Lukaszewiez, J., Perkal, J., Steinhaus, H., & Zubrzchi, S. (1951). Sur la liason: Division des points d'un ensemble fini. *Colloquium Mathematicum, 2*, 282–285.

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition, 1*(2), 131-149.

Landau, S. & Chis Ster, I. (2010). Cluster Analysis: Overview. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education, 3rd edition* (pp. 72-83). Oxford, UK: Elsevier Ltd.

Mazzocchi, M. (2008). *Statistics for Marketing and Consumer Research*. London, UK: Sage Publications Ltd.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15-24.

Norusis, M. J. (2010). Chapter 16: Cluster analysis. *PASW Statistics 18 Statistical Procedures Companion* (pp. 361-391). Upper Saddle River, NJ: Prentice Hall.

SAS Institute (1983). *SAS technical report A-108 Cubic Clustering Criterion*. Cary, NC: SAS Institute Inc. Retrieved from https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf

Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology, 17*, 201–226.

Sokal R. R. & Michener C. D. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin, 38,* 1409-1438.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of*

numerical taxonomy. San Francisco: W. H. Freeman.

Wilkinson, L., Engelman, L., Corter, J., & Coward, M. (2000). Cluster analysis. In L. Wilkinson (Ed.), *Systat 10 – Statistics I* (pp. 65-124). Chicago, IL: SPSS Inc.

Wilmink, F. W. & Uytterschaut, H. T. (1984). Cluster analysis, history, theory and applications. In G. N. van Vark & W.W. Howells (Eds.), *Multivariate Statistical Methods in Physical Anthropology* (pp. 135-175). Dordrecht, The Netherlands: D. Reidel Publishing Company.

Yim, O. & Bialystok, E. (2012). Degree of conversational code-switching enhances verbal task switching in Cantonese-English bilinguals. *Bilingualism: Language and Cognition, 15*(4), 873-883.

Appendix: SPSS Syntax for Hierarchical Cluster Analysis

The steps outlined in the tutorial can be performed using the following syntax in SPSS. The entry 'C:\Users\cluster.tmp' is the location of a temporary file for the analysis (suitable for both PC and Mac operating systems). The user can substitute this line with a specific location on their computer if they wish and it can be deleted after the analysis is complete. The comment lines beginning with an asterisk can be copied into the SPSS syntax window for reference.

```
*PROXIMITIES: Substitute with your own variable names in your datafile (as many as desired).
*MEASURE: Distance measure (e.g., Squared Euclidean).
*STANDARDIZE: Transformation applied to the variables (e.g., Range -1 to 1).
*METHOD: Linkage measure (e.g., between-groups average; others are SINGLE or COMPLETE).
*To generate a cluster variable in the Data View window, add the following line under
* the CLUSTER command: /SAVE CLUSTER (number of clusters desired).

PROXIMITIES Variable1 Variable2 Variable3
  /MATRIX OUT ('C:\Users\cluster.tmp')
  /VIEW=CASE
  /MEASURE=SEUCLID
  /PRINT NONE
  /STANDARDIZE=VARIABLE RANGE.

CLUSTER
  /MATRIX IN ('C:\Users\cluster.tmp')
  /METHOD BAVERAGE
  /PRINT SCHEDULE
  /PRINT DISTANCE
  /PLOT DENDROGRAM VICICLE.
```

Citation

Yim, O., & Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology*, *11* (1), 8-21.

Received: 16/08/14 ~ Accepted: 22/10/14