Segmentation Using Two-Step Cluster Analysis

Aaron Tkaczynski

Abstract The purpose of this chapter is to explain the rationale for employing TwoStep cluster analysis as a market segmentation method within social marketing. Here, the key stages to be performed and the validation techniques required for effective application of this clustering technique are outlined. To further support the application of this cluster analysis technique as a profiling tool, a review of 25 recent market segmentation studies that have utilised this method is provided. Finally, a case study is provided to demonstrate how TwoStep cluster analysis is employed to segment respondents for an active school travel social marketing campaign that was being developed in Queensland at time of writing. Based on a sample of 537 respondents, three segments were identified and validated, each of which differed significantly based on psychographic, behaviour, geographic and demographic variables. Limitations of the TwoStep Cluster Analysis method are also provided, and opportunities for future research employing TwoStep cluster analysis within a social marketing context conclude this chapter.

Introduction

TwoStep Cluster Analysis is a cluster analysis algorithm that is available in Predictive Analytics SoftWare (PASW). TwoStep Cluster Analysis is a statistical procedure that is employed by a user to identify similar groups or "clusters" of people or objects within data sets (Norusis 2011). This segmentation method allows users to retain full information, providing rich explanation for managerial decision-making purposes. TwoStep cluster analysis is also considered more reliable and accurate when compared to traditional clustering methods such as the k-means clustering algorithm (Norusis 2007). Since being introduced in Version 11.5 of the Statistical Packages for the Social Sciences (SPSS), TwoStep cluster analysis has been increasingly utilised in a variety of fields such as tourism (Hsu

A. Tkaczynski (✉)
University of Queensland, Brisbane, Australia
e-mail: a.tkaczynski@uq.edu.au

T. Dietrich et al. (eds.), *Segmentation in Social Marketing*,
DOI 10.1007/978-981-10-1835-0_8

et al. 2006; Tkaczynski et al. 2015), health (Griffin et al. 2014; McLernon et al. 2012), transport (Cerin et al. 2007; Chang and Yeh 2007) and psychology (Fillman et al. 2013; Ulstein et al. 2007).

TwoStep Clustering Procedure

The TwoStep clustering procedure, as the name suggests, involves two distinct stages. As a first phase, original cases are grouped into preclusters (Okazaki 2007). The goal of this step, classed as *preclustering*, is to reduce the size of the matrix that contains distances between all possible pairs of cases (Norusis 2011). This clustering method assumes that all variables are independent, that continuous variables have a normal distribution, and categorical variables have a multinominal distribution (Norusis 2007). The categorical and ordinal variables are treated as nominal. The cluster parameter employs a hierarchical method and the scale parameter for each continuous variable is the standard deviation of each continuous variable. If this is unavailable, the default is one (IBM 2011). If categorical and continuous variables are employed, the log-likelihood algorithm is required. Alternatively, if only continuous items are to be analysed, the Euclidian algorithm can be chosen. Based on this procedure, it is assumed that the variances are identical over variables and clusters. The cases represent the objects to be clustered, whereas the variables represent attributes on which clustering is based (Norusis 2007). The algorithm randomly chooses to assign an observed case to a cluster. As each case is read, the algorithm decides if the current case should be merged with a previously formed precluster. Alternatively, the algorithm can choose to start a new precluster. When preclustering is complete, all cases in the same precluster are treated as a single entity (Norusis 2007).

In the second step, the preclusters are clustered using the hierarchical clustering algorithm. This stage is classed as *clustering*. Forming clusters hierarchically lets the researcher explore a range of solutions with different numbers of clusters (Norusis 2007). This stage produces a range of solutions which is then reduced to the best number of clusters based on the Schwarz's Bayesian information criterion (BIC) (Norusis 2011). In addition, outliers can be identified and screened out in the algorithm (Chiu et al. 2001). Once the cluster solution is formed, chi-square tests are conducted for categorical variables and student t-tests for continuous variables to examine the importance of individual variables in a cluster and to identify if the item is valid in the total solution (Norusis 2007). If an item has an insignificant value ($p > 0.05$) it is invalid and should be removed from the analysis. The TwoStep cluster analysis is then rerun until only valid items remain.

Whilst TwoStep cluster analysis can be completed within two stages, an additional phase that is available for market segmentation researchers is running chi-square tests for binary or dichotomous variables once the clusters have been formed. TwoStep cluster analysis creates a cluster membership variable that allows variables that may have been combined into one to be tested for their significance at

the individual level. This helps to overcome the limitation of one type of variable (e.g. motivations, interests) biasing a cluster solution to a specific type of variable/s. TwoStep cluster analysis treats all individual variables with equal importance. If 20 motivational items, four demographic items and a geographic variable were analysed simultaneously, this would bias the cluster solution to a motivation focus due to 80 % of the items relating to this type of variable. Therefore, for variables that contain multiple items (Rundle-Thiele et al. 2015; Tkaczynski and Prebensen 2012) which were relevant in forming clusters in the TwoStep cluster analysis, can be individually tested (e.g. number of activities). This determines whether each segment is significantly different from the other items based on the individual items.

TwoStep Cluster Validation

Four final validation techniques need to be employed for a TwoStep cluster analysis solution to be accepted. First, when using the BIC for statistical inference, the silhouette measure of cohesion and separation is required to be at or above the required level of 0.0 (highest being 1.0). This stage measures the relationship of the variables within and between clusters. A score above 0.0 would ensure that the within-cluster distance and the between-cluster distance was valid among the different variables as there is some variation between variables (Norusis 2011). It is more beneficial if the silhouette measure of cohesion and separation is above 0.2 as this showcases that there is a *fair* separation distance between clusters.

Second, all variables within a solution need to be statistically significant ($p < 0.05$). Thus, insignificant variables should be removed from the analysis. Consequently, variables that might be particularly relevant to the study (e.g. gender, club membership) might need to be removed if there is no difference between the clusters based on this variable. Recall, that for market segmentation to be purposeful, clusters need to be distinguished on different classifying variables. The removed variables might still be important for social marketing strategies, but they are irrelevant for differentiating respondents.

Third, when considering the input (predictor) importance to determine the importance of variables in a cluster solution, variables with a low rating (0.02 or below) must be carefully considered for their usage in the final solution. Items with a negative value should be removed from the analysis due to being insignificant (Tkaczynski et al. 2015). Usually, these variables will be the same variables that are outlined in the second validation stage—i.e. those that are statistically insignificant. Variables that have a predictive importance of 0.00 or 0.01 can be included, but it should be noted that the responses to these variables will likely be similar across the different clusters.

The fourth and final validation technique that is recommended by multivariate analysis experts (e.g. Hair et al. 2006) is to randomly split the sample in two and compare the results with the final solution. If the same number of clusters is found in both the final and split solutions, and the characteristics and significance

variables of the solutions are similar, then validation is confirmed. Note, for a cluster solution to have a higher chance of being validated, it is recommended that the user collects a large sample size. As segmentation authors such as Dolnicar et al. (2014) argue, for valid results, at least 70 cases should be employed for each variable in data driven segmentation research; therefore, a small sample size (e.g. <300) would be very difficult to validate with only 150 cases in each split solution, particularly if many variables are being cluster analysed simultaneously.

TwoStep Cluster Analysis: An Alternative Solution

While a variety of statistical techniques such as exploratory factor analysis, Pearson's chi-square test, bootstrap analysis and k-means clustering deliver beneficial findings to market segmentation researchers, TwoStep cluster analysis enables an alternative approach to market segmentation which provides distinct advantages, as outlined below.

Application of both categorical and continuous data The first and arguably the greatest advantage of TwoStep cluster analysis is its ability to segment data based on any form of data measurement (e.g. binary, Likert or categorical) simultaneously. Thus, while certain forms of analysis such as k-means clustering require numeric measurement to work effectively, the TwoStep cluster analysis algorithm standardises all of the variables unless the option is specifically overridden by the user (Norusis 2011). As distance measures can be quite sensitive to differing scales or magnitudes among the variables, and variables with larger dispersion (e.g. larger standard deviations) have more impact on the final similarity value (Hair et al. 2006), it has been argued that clustering variables should be standardised whenever possible (Baeza-Yates 1992). TwoStep cluster analysis not only ensures that through standardisation, one variable does not dominate the cluster solution, but also that these variables can be overridden if required (Norusis 2011).

Works well with large data sets The second benefit of TwoStep cluster analysis is that it works extremely well on large data sets. While this clustering method has been employed in social marketing contexts (e.g. Stranak et al. 2014; Ulstein et al. 2007) with small data sets ($n < 500$), it is more advantageous when applied to large data sets (Norusis 2011). Large data sets available, such as census data from major industries such as health, finance, religion or education, can be utilised by researchers with great success (e.g. Rundle-Thiele et al. 2015; Griffin et al. 2014).

Automatically determines the number of clusters A third advantage of TwoStep cluster analysis is that unless specifically overridden by the user, the cluster algorithm automatically determines the number of clusters within a cluster solution (Tkaczynski et al. 2010). Therefore, if the research is exploratory and the characteristics of groups are not known a priori, TwoStep cluster analysis provides a viable solution to a user for determining how many clusters (groups) might be within the data. As a consequence, the user's judgment is not the determining factor when identifying the number of clusters, which can be very beneficial when trying

to identify constructs of clusters and the most significant and relevant segmentation variables.

Determines the predictor importance of variables in a cluster solution The fourth advantage of TwoStep cluster analysis is that it enables the user to identify the importance of each item in the cluster solution and how it might be statistically significantly different amongst clusters post analysis. This can be very important when seeking to determine how relevant a specific variable is to the total solution (Tkaczynski et al. 2015). For example, while social marketing studies will traditionally use psychographic (e.g. motivations, perceptions, interests) or behavioural (e.g. physical activity participation, media usage, club membership) variables as a first phase of classifying segments (i.e. people) into groups, the importance of these variables in differentiating the cluster solutions might be minimal, or even insignificant. Rather, it is descriptive variables (e.g. age and gender) that are often used to distinguish psychographic or behavioural items as a post analysis validity measure (e.g. Atlantis et al. 2009; Dietrich et al. 2015a) which might provide the most differentiation in a cluster solution. Identifying which variables are most important in a cluster solution can help market segmentation researchers to plan for differences by focusing on these key variable differences when applying strategic marketing plans or market communication strategies.

Social Marketing and TwoStep Cluster Analysis: A Review

To further examine TwoStep Cluster Analysis's potential for social marketing, a review (see Table 1) is undertaken. These 25 studies were conducted in a variety of countries, such as Australia (Atlantis et al. 2009; Dietrich et al. 2015a), US (e.g. Hu et al. 2009; Stranak et al. 2014) and Norway (Glasø et al. 2007; Ulstein et al. 2007). The focus of the studies varies immensely but a common emphasis on health practices for social good, such as healthy food consumption (Honkanen 2010; Hu et al. 2009), alleviating stress and anxiety (Nielsen & Knardahl 2014; Ulstein et al. 2007), and moderating substance usage (Fleury et al. 2015; Mason and Korpela 2009) was identified. There is a major focus on differentiating patients currently struggling with identified social issues (Bamvita et al. 2014; Créton et al. 2009) based on predefined variables such as demographics and geographic region lived.

In examining the applicability of TwoStep cluster analysis to social marketing, it is noted that a high percentage (68 %) of studies have a sample size smaller than 500. Therefore, one of the most pronounced and promoted strengths of TwoStep cluster analysis—working well with large data sets (Norusis 2011)—is not being fully utilised within these market segmentation studies. Almost two-thirds (64 %) of the studies develop three or four valid clusters (segments), and have a main focus of aiming to differentiate respondents on a pre-conceived classification variable such as behaviour (Chan et al. 2005; Rompré et al. 2007). Despite TwoStep cluster analysis's ability to develop significant and valid clusters as a sole segmentation method, nearly all of the studies employ additional methods pre or post cluster

Table 1 Social marketing studies

Author	Country	Study focus	Respondents	Sample size	Number of clusters	Other methods
Atlantis et al. (2009)	Australia	Metabolism	Male patients	1195	2	II, IV
Bamvinia et al. (2014)	Canada	Hepatitis C virus	Patients	60	4	II
Chan et al. (2005)	Hong Kong	Computerisation skills	Practitioners	954	3	II, VIII, IX
Chan et al. (2006)	Hong Kong	Spiritual care	Part-time nurses	193	3	I, II, V
Créton et al. (2009)	Netherlands	Hypodontia characteristics	Patients	189	4	IV
Dietrich et al. (2015b)	Australia	Alcohol consumption	High school students	2114	3	I, II, IV, IX
Dietrich et al. (2015a)	Australia	Alcohol consumption	High school students	371	3	VII
Fairburn et al. (2007)	England	Eating disorders	Patients	170	4	IV, VIII
Ferreira et al. (2008)	Brazil	Quality of life	Cancer patients	113	2	I, III, IV, VIII
Fleury et al. (2015)	Canada	Substance dependence	Participants	121	4	III
Glasø et al. (2007)	Norway	Workplace bullying	Victims, Non victims	144	2	I, IV, VI, VII
Griffin et al. (2014)	Australia	Health behaviour	Older Australians	96276	6	I, III, VII, IX
Helm and Eis (2007)	Germany	Chemical susceptibility	Outpatients	196	3	II, VIII
Honkanen (2010)	Russia	Food preferences	Consumers	1081	5	I, II, VII
Hu et al. (2009)	America	Blueberry jam attributes	Customers	202	2	I, IX
Lopez-Alonzo et al. (2014)	Spain	Motor evoked potentials	Respondents	56	2	I, II, IV, VII
McLernon et al. (2012)	Scotland	Lifestyle choices	Older women	3218	3	II, IX
Mason and Korpella (2009)	America	Substance use and health	Adolescents	68	2	I
Murphy and Marelich (2008)	America	Young children resiliency	Children	111	2	I, IX
Nielsen and Knardahl (2014)	Norway	Coping strategies	Employees	3738	3	I, IV, VI, VII

(continued)

Table 1 (continued)

Author	Country	Study focus	Respondents	Sample size	Number of clusters	Other methods
Polymeros et al. (2015)	Greece	Consumer preferences	Consumers	149	2	II
Rompré et al. (2007)	Canada	Sleep bruxism	Participants	143	3	I, VII
Rundle-Thiele et al. (2015)	Australia	Physical activity	Residents	1459	4	
Stranak et al. (2014)	America	Hypotension strategies	Physicians	216	4	I, II
Ulstein et al. (2007)	Norway	Relative stress scale	Carers	194	3	I, VII

Note I = Descriptive statistics, II = chi-square test, III = regression, IV = t-test, V = factor analysis, VI = correlation, VII = ANOVA, VIII = Mann Whitney, IX = Other

analysis to further explain respondents. Not surprisingly, descriptive statistics (56 %) is the most popular method, followed by chi-square (36 %) and analysis of variance (32 %).

Social Marketing Case Study

To further showcase the strength and applicability of TwoStep cluster analysis, this clustering technique was employed within a larger social marketing formative study involving carers of primary school-age children in Queensland, Australia. This study aimed to explore the behaviours, behavioural intentions, attitudes, social norms and perceived risks in the context of active school travel for primary school-aged children.

An online questionnaire containing 32 items was designed for this project to cover four bases of segmentation (Kotler and Armstrong 2008). The questionnaire included a series of demographic (i.e. age [of child], carer's age, gender, gender [of child], carer's relationship [with child], education level, weekly income, number of cars, responsibility for getting the child/children to/from school, employment status, geographic (place of residence, distance from school), psychographic (three intention items, three perceived risk items, three social norm items, two perceived behavioural control items, eight attitude items) and behavioural (transport mode) questions. The questionnaire was completed by parents (carers) of primary school-aged children across a variety of regions throughout Queensland including Brisbane, the Gold Coast and the Wide Bay-Burnett region. To increase participation, incentives of winning one of ten AUD$30 gift vouchers were offered.

In total, 537 respondents completed the survey and these responses were analysed using TwoStep Cluster Analysis (Version 22.0). This method was specifically

chosen for the analysis since (1) both continuous (e.g. attitudes, risk perceptions) and categorical (e.g. child gender, distance from school) measures were used in the study, (2) an available large sample size (n > 500) would allow TwoStep Cluster Analysis to produce potentially reliable and valid segments based on several key classification variables, and (3) this was an exploratory study in which the number of clusters could not be determined in advance and the user allowed TwoStep Cluster Analysis to automatically determine the number of clusters.

TwoStep Cluster Validation

Recall that variables within solutions need to be identified as a requirement for cluster validation. While the inceptive cluster analysis initially produced three clusters, the solution could not be validated. Most noticeably, *your relationship to the child* and *your gender* items were not significant and generated limited predictor importance in the cluster solution—most respondents were *mothers* (over 90 %). Respondents were also largely aged between *30 and 40, employed* and had a family weekly income in excess of *AUS$2000 per week*. Furthermore, the majority of children in the study were aged *under 6* and respondents were not differentiated by the region [in Queensland] in which they lived. As these items were all insignificant ($p > 0.05$) they were removed from further analysis.

The cluster analysis was rerun and a three cluster solution was again formed. All 25 items were identified as significant ($p < 0.05$) and contributed to predictive importance in cluster formation. When the file was split in two for validation purposes, it was also confirmed that the same number of clusters could be identified in both the split solutions and the respondent characteristics and predictive importance of the variables for the three clusters was similar to the final solution. Consequently, the TwoStep cluster analysis solution was confirmed for this study.

Examining the Segments Generated by TwoStep Cluster Analysis

The average silhouette measure of cohesion and separation was 0.3 for the cluster solution. This indicates that the distance measured between clusters was *fair* and therefore acceptable for further analysis. Tables 2 and 3 describe the clusters. Table 2 consists of the continuous variables which were all measured on a bi-polar (−3 strongly disagree to +3 strongly agree) scale. Table 3 outlines the categorical variables. The predictive importance of all variables in the TwoStep cluster analysis is listed in brackets next to each variable. As mentioned previously, if an item has a rating of between 0.8 and 1.0, it is extremely important in predicting cluster formation. Conversely, items with a score of 0.0–0.2, while significant, are less important in forming the three clusters.

Table 2 Cluster Solution (continuous variables)

Variable	Non-walking oriented parents—n = 62 (13.0 %)	Long distance safety concerned parents—n = 283 (59.5 %)	Walking-focused health conscious parents—n = 131 (27.5 %)
Attitude			
Walking to/from school is good/bad (1.00)	−2.47	1.82	2.61
Walking to/from school is valuable/worthless (0.92)	−2.42	1.69	2.47
Walking to/from school is beneficial/harmful (0.82)	−2.33	1.84	2.62
Walking to/from school is enjoyable/unenjoyable (0.63)	−2.13	1.35	1.95
Walking to/from school is healthy/unhealthy (0.60)	−0.47	2.73	2.98
Walking to/from school is pleasant/unpleasant (0.55)	−2.06	1.24	1.82
Walking to/from school is exciting/boring (0.43)	−1.79	1.01	1.56
Walking to/from school is important/unimportant (0.35)	−1.10	1.17	2.27
Intentions			
I plan to increase the number of times the child walks to/from school this week (0.48)	−2.31	−2.88	−0.53
I will increase the number of times the child walks to school this week (0.48)	−2.32	−2.88	−0.53
I intent to increase the number of times the child walks to/from school this week (0.43)	−2.23	−2.33	−0.5
Perceived Behavioural Control			
The distance between the school and the child's home is too far to walk (0.69)	0.48	2.11	−2.17
How much do you feel that the child walking to/from school next week is beyond your control? (0.03)	0.65	0.37	1.37
Perceived risk			
The traffic along the route to/from school makes the walk unsafe (0.24)	0.74	1.89	−0.27
Streets are dangerous to cross along the route to/from school (0.18)	0.89	1.93	0.08

(continued)

Table 2 (continued)

Variable	Non-walking oriented parents—n = 62 (13.0 %)	Long distance safety concerned parents—n = 283 (59.5 %)	Walking-focused health conscious parents—n = 131 (27.5 %)
The dangers of crime along the route to/from school makes the walk unsafe (0.14)	−0.48	0.22	−1.66
Social norms			
People who are important to me walk their children to/from school (0.36)	−1.23	−1.51	0.98
People who are important to me think the child should/should not walk to/from school (0.29)	−1.06	−1.24	1.02
People who are important to me would disapprove/approve of me walking my child to school (0.20)	−1.05	−0.96	1.08

Note the number in brackets after the variable represents the importance of the variable in cluster formation. This is between 1.0 and 0.0. The closer to 1.0, the more important it is

From viewing Tables 2 and 3 in Chap. 7, it can be noted that the attitude variables, *walking to school is good/bad* (1.00), *valuable/worthless* (0.92) and *beneficial/harmful* (0.82), have the highest predictive importance amongst all variables and are the most relevant in defining differences amongst the three clusters. Furthermore, *distance between the school and the child's home is too far to walk* (0.69), *transport mode* (0.52) and *distance from child's home to school* (0.47) also have relatively high predictive importance. Variables of less importance to cluster formation include *gender of child* (0.02), *child responsibility* (0.02), and *education level* (0.01).

Non-walking Oriented Parents

In seeking to understand the three clusters, the following exploratory notes are provided. The first cluster is the smallest (13.0 %) and has a negative attitude towards their child/children walking to school. Responses such as *walking is bad* (−2.47), *worthless* (−2.42) and *harmful* (−2.33) are identified. Despite a high percentage of respondents (32.3 %) living close to the school (<2 km), these guardians largely drive their children to school in a *family vehicle* (48.4 %), and appear to care less how they are perceived by friends and family when considering the form of transport that their children used when going to and from school. This cluster is the least educated (45.7 % did not have a university degree) and is the

Table 3 Cluster solution (categorical variables)

Variable	Non-walking oriented parents—n = 62 (13.0 %)	Long distance safety concerned parents—n = 283 (59.5 %)	Walking-focused health conscious parents—n = 131 (27.5 %)
Transport mode (0.52)			
Walk	4.8	0.0	23.7
Bicycle	8.1	0.7	3.1
Family vehicle	48.4	78.4	8.4
Carpool	1.6	1.1	0.8
Bus	6.5	3.5	10.7
Walk + Bicycle + Family vehicle	1.6	0.0	9.2
Walk + Family vehicle	11.3	0.4	35.1
Family vehicle + Carpool	6.5	1.8	1.5
Family vehicle + Bus	3.2	6.4	0.0
Two or more (not already chosen)	4.8	3.5	8.4
Three or more (not already chosen)	3.2	1.8	9.9
Other	0.0	2.5	0.0
What is the approximate distance from the child's home to school? (0.47)			
<1 km	7.8	2.1	40.5
1–2 km	32.3	6.4	44.3
2–3 km	9.7	15.2	10.7
3–4 km	8.1	13.4	2.3
4–5 km	14.5	8.8	0.8
5 km+	27.4	54.1	1.5
Number of cars (0.03)			
None	0.0	14.1	2.3
1 car	24.2	21.9	41.9
2 cars	61.3	66.4	44.3
>2 cars	14.5	10.2	11.5
Are you responsible for getting the child/children to/from school? (0.02)			
Yes	71.0	86.6	76.3
Sometimes	27.4	11.3	22.9
No	1.6	2.1	0.8
Gender of child (0.02)			
Male	56.5	45.9	59.5
Female	43.5	54.1	40.6
Education level of respondent (0.01)			
School	12.9	15.9	10.7
Diploma	41.9	28.6	26.7

(continued)

Table 3 (continued)

Variable	Non-walking oriented parents—n = 62 (13.0 %)	Long distance safety concerned parents—n = 283 (59.5 %)	Walking-focused health conscious parents—n = 131 (27.5 %)
Bachelor degree	27.4	40.3	35.9
Postgraduate degree	17.7	15.2	26.7

Note the number in brackets after the variable represents the importance of the variable in cluster formation. This is between 1.0 and 0.0. The closer to 1.0, the more important it is

least responsible (71.0 %) for how their children arrive and depart school. Based on these key defining characteristics, this cluster is defined as *non-walking orientated parents*.

Long Distance Safety Concerned Parents

Cluster two, which is the largest cluster (59.5 %), has a positive attitude towards their children walking to school. They rate *walking to school is healthy/unhealthy* particularly high (2.73). Despite its positive attitude towards this form of physical exercise, this cluster is extremely negative (e.g. −2.88 for two items) in the likelihood of their child/children walking more to or from school in the next week. Over half (54.1 %) of these respondents live *over 5kms from their child/children's school* and the child/children under their supervision were *girls* (most likely under the age of 6). This cluster rates the risk items of walking to school most highly amongst clusters and also *drive* their child/children to school most frequently in a family vehicle (78.4 %). Of interest is that a very small percentage of this segment allows their children to also arrive/depart school via *bicycle* (0.7 %), *carpool* (1.1 %) or a *bus* (3.5 %). These respondents are also most responsible (86.6 %) for how their child/children get to and from school, and do not appear to care how they are perceived by their friends or family for not allowing their children to walk to school, with negative ratings for all of the perception items. Based on the key defining cluster characteristics, this segment is defined as *long distance safety concerned parents*.

Walking-Focused Health Conscious Parents

The third cluster represents approximately a quarter (27.5 %) of respondents. These respondents have the most positive attitude towards their children walking to/from school out of the clusters. For example, cluster three respondents provide almost a perfect score (2.98) for considering walking to school as *healthy*. This cluster is

distinguishable based on the majority of respondents (84.8 %) living within a proximal distance (< 2 km) to their child's school. Despite cluster three respondents having *two cars*, these respondents have a high percentage of their children walking to school (23.7 %) and noticeably, rated the risks of their child/children walking to school as the lowest amongst clusters. This is likely due to parents knowing the paths the child/children might take based on the proximity of the school to their house. This cluster agrees that impressing people with allowing their children to walk to and from school is important (e.g. people who are important to me would approve of me walking my child to school =1.08), and had the highest control (1.37) of whether their child/children walked to school. This cluster was also the most educated (62.6 % had finished a university degree) and also had the highest percentage of children being male (59.5 %). Due to the focus on their children walking to school and the positive perceptions and attitudes that this activity exhibits to these respondents, this cluster is labelled as *walking-focused health conscious parents.*

Discussion and Conclusion

It is argued by Lefebvre (2013, p. 125) that the *"core of social marketing is the people we intend to serve"* and that "*segmentation reinforces and builds on the core tenet of marketing that we should be customer or people focused.*" By segmenting people based on key social marketing criteria, such as their attitude towards physical activity and substance consumption, marketers can design messages, products and services to potentially enable people to engage in positive behavioural changes which will improve their lifestyle and, potentially, their psychological, emotional and physical well-being. This chapter provides (1) an outline of the TwoStep cluster analysis procedure, (2) a review of TwoStep cluster analysis studies conducted in a social marketing context, and (3) a case study to demonstrate use of TwoStep cluster analysis in social marketing. Future considerations based on the active school travel case study are discussed below.

For market segmentation to be relevant, it needs to purposeful. In other words, segments need to be accessible, actionable, sustainable and measurable (Kotler and Armstrong 2008). Many social marketing practitioners work for not-for-profit institutions and segmentation can be a viable tool to target the most fruitful segments (i.e. readiness to change). Examining the walking to school segmentation results, it can be argued that the *long distance safety concerned parents* (cluster 2) need to be a priority segment for Queensland schools and health practitioners. This segment represents over half of the sample, so investment within this group of people could potentially provide positive manifested outcomes if targeted appropriately. Notwithstanding their positive attitude towards walking, parents within this cluster live a long distance from their children's schools (over 5 km), which presents a geographical barrier that cannot be easily overcome. It is suggested that providing infrastructure tools to help realise this cluster's walking to school

intention is the key. Another recommended option is to include designated drop-off zones and walking school buses in combination as a solution to deliver safe active travel options for this segment.

The third cluster, *walking-focused health conscious parents*, is a prime segment that should be targeted. These respondents live within a close radius of their child's school across the different regions of Queensland, and have their children actively walking to and from school. However, as less than a quarter (23.7 %) choose this as the only method, additional marketing of the benefits of this activity could be an option. However that external factors such as rain or time restrictions (e.g. running late) might imply (but is not confirmed in this study) that a family vehicle is sometimes required, further promoting the health, emotional and physical benefits of walking could increase the numbers of children walking as their principal transport mode. Furthermore, parents could act as group leaders for walking groups of students or organise activities such as "ride your bike to school day" to encourage more active to and from school travel.

The *non-walking orientated parents* cluster has a negative attitude towards walking and is thus a challenging segment. However, given this segment is small and is likely hard to please, it should not be targeted specifically. Recall that for segmentation to be purposeful, segments need to be accessible, actionable, sustainable and measurable. Since this segment is the least responsible for their child getting to and from school, it is hoped that when their children are older (perhaps 9 or 10) they may, based on the promoted benefits of daily walking, choose to do so for themselves.

Limitations and Opportunities for Future Research

TwoStep cluster analysis provides many benefits for social marketers. A major limitation of the method has been that in previous versions (before 20.0) of TwoStep cluster solution, the user had the choice to consider a solution with or without missing data. However, the option with missing data will usually present a higher BIC and will consequently be preferred. Ultimately, TwoStep cluster analysis now removes all cases with missing data. Although this study had limited missing data, other studies with items that are known to have a high number of non-responses (such as age or annual income) could have a high percentage of their sample removed. A recommendation here is to design questionnaires in such a way that options for non-response are limited, for example, compulsory online survey options or providing a financial motivation to complete paper surveys. A second major limitation is that TwoStep cluster analysis is extremely sensitive to changes in entry, and the final solution may depend on the order of the cases in the file. Therefore, in splitting the file in two, it is recommended that every odd and even case be considered, since time differences or regional differences could produce quite divergent results when cases are split. A third limitation of TwoStep cluster analysis is that while certain academic papers have successfully employed this

method to profile customers, more research is required to compare and contrast the different clustering methods (e.g. TwoStep, k-means, R). Consequently, the strength of TwoStep cluster analysis in comparison to other methods is relatively unknown. There is an opportunity for future statistical research to improve this useful method by comparing its strengths and weaknesses with other clustering methods.

References

Atlantis, E., Martin, S. A., Haren, M. T., Taylor, A. W., & Witter, T. G. A. (2009). Inverse associations between muscle mass, strength, and the metabolic syndrome. *Metabolism, Clinical and Experimental, 58*, 1013–1022.

Baeza-Yates, R. A. (1992). Introduction to data structures and algorithms related to information retrieval. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: data structures and algorithms*. Upper Saddle River, NJ: Prentice Hall.

Bamvita, J. M., Roy, E., Zang, G., Jutras-Aswad, D., Artenie, A. A., Levesque, A., et al. (2014). Portraying persons who inject drugs recently infected with hepatitis C accessing antiviral treatment: A cluster analysis. *Hepatitis Research and Treatment*, 1–7.

Cerin, E., Leslie, E., Du Toit, L., Owen, N., & Frank, L. D. (2007). Destinations that matter: Associations with walking for transport. *Health & Place, 13*, 713–724.

Chan, M. F., Chung, L., Y. F., Lee, A. S. C., Wong, W. K., Lee, G. S. C., Lau, C. Y., et al. (2006). Investigating spiritual care perceptions and practice patterns in Hong Kong nurses: Results of a cluster analysis. *Nurse Education Today, 26*, 139–150.

Chan, M. F., Day, M. C., Suen, L. K. P., Tse, S. H. M., & Tong, T. F. (2005). Attitudes and skills of Hong Kong Chinese medicine practitioners towards computerization in practice: A cluster analysis. *Medical Informatics and the Internet in Medicine, 30*, 55–68.

Chang, H. L., & Yeh, T. H. (2007). Motorcyclist accident involvement by age, gender, and risky behaviors in Taipei, Taiwan. *Transportation Research Part F: Traffic Psychology and Behaviour, 10*, 109–122.

Chiu, T., Fang, D. P., Chen, J., & Wang, Y. J. C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *7th ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: ACM SIGKDDD.

Creton, M., Cune, M. S., De Putter, C., Ruijter, M., & Kuijpers-Jagtam, A. M. (2009). Dentofacial characteristics of patients with hypodontia. *Clinical Oral Investigations, 14*, 467–477.

Dietrich, T., Rundle-Thiele, S., Leo, C., & Connor, J. P. (2015a). One size (Never) fits all: Segment differences observed following a school-based alcohol social marketing program. *Journal of School Health, 85*, 251–259.

Dietrich, T., Rundle-Thiele, S., Schuster, L., Drennan, J., Russell-Bennett, R., Leo, C., et al. (2015b). Differential segmentation responses to an alcohol social marketing program. *Addictive Behaviors, 49*, 68–77.

Dolnicar, S., Grun, B., Leisch, F., & Schmidt, K. (2014). Required sample sizes for data-driven market segmentation analyses in tourism. *Journal of Travel Research, 53*, 296–306.

Fairburn, C. G., Cooper, Z., Bohn, K., O'Connor, M. E., Doll, H. A., & Palmer, R. L. (2007). The severity and status of eating disorder NOS: Implications for DSM-V. *Behaviour Research and Therapy, 45*, 1705–1715.

Ferreira, K. A. S. L., Kimura, M., Teixeira, M. J., Mendoza, T. R., Da NóBrega, J. C. M., Graziani, S. R., et al. (2008). Impact of cancer-related symptom synergisms on health-related quality of life and performance status. *Journal of Pain and Symptom Management, 35*, 604–616.

Fillman, S. G., Cloonan, N., Catts, V. S., Miller, L. C., Wong, J., McCrossin, T., et al. (2013). Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Molecular Psychiatry, 18*, 206–214.

Fleury, M. J., Grenier, G., Bamvita, J. M., Perreault, M., & Caron, J. (2015). Typology of individuals with substance dependence based on a montreal longitudinal catchment area study. *Administration and Policy in Mental Health and Mental Health Services Research, 42*, 405–419.

Glaso, L., Matthiesen, S. B., Nielsen, M. B., & Einarsen, S. (2007). Do targets of workplace bullying portray a general victim personality profile? *Scandinavian Journal of Psychology, 48*, 313–319.

Griffin, B., Sherman, K. A., Jones, M., & Bayl-Smith, P. (2014). The clustering of health behaviours in older Australians and its association with physical and psychological status, and sociodemographic indicators. *Annals of Behavioral Science, 42*, 205–214.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall.

Helm, D., & Eis, D. (2007). Subgrouping outpatients of an environmental medicine unit using SCL-90-R and cluster analysis. *International Journal of Hygiene and Environmental Health, 210*, 701–713.

Honkanen, P. (2010). Food preference based segments in Russia. *Food Quality and Preference, 21*, 65–74.

Hsu, C. H. C., Kang, S. K., & LAM, T. (2006). Reference group influences among Chinese travelers. *Journal of Travel Research, 44,* 474–484.

Hu, W., Woods, T., & Bastin, S. (2009). Consumer Cluster analysis and demand for blueberry jam attributes. *Journal of Food Products Marketing, 15*, 420–435.

IBM. (2011). *Two step Cluster analysis* [Online]. IBM SPSS Statistics Information Center. http://pic.dhe.ibm.com/infocenter/spssstat/v21r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fclusterviewer_clusters_panel.htm

Kotler, P., & Armstrong, G. M. (2008). *Principles of marketing*. California, USA: Prentice-Hall.

Lefebvre, R. C. (2013). *Social marketing and social change*. USA: Jossey-Bass.

Lopez-Alonzo, V., Cheeran, B., Rio-Rodriguez, D., & Fernandez-Del-Olmo, M. (2014). Inter-individual variability in response to non-invasive brain stimulation paradigms. *Brain Stimulation, 7*, 372–380.

Mason, M. J., & Korpela, K. (2009). Activity spaces and urban adolescent substance use and emotional health. *Journal of Adolescence, 32*, 925–939.

Mclernon, D. J., Powell, J. J., Jugdaohsingh, R., & Macdonald, H. M. (2012). Do lifestyle choices explain the effect of alcohol on bone mineral density in women around menopause? *The American Journal of Clinical Nutrition, 95*, 1261–1269.

Murphy, D. A., & Marelich, W. D. (2008). Resiliency in young children whose mothers are living with HIV/AIDS. *AIDS Care, 20*, 284–291.

Nielsen, M. B., & Knardahl, S. (2014). Coping strategies: A prospective study of patterns, stability, and relationships with psychological distress. *Scandinavian Journal of Psychology, 55*, 142–150.

Norusis, M. J. (2007). *SPSS 15.0 advanced statistical procedures companion.* Chicago, IL: Prentice Hall.

Norusis, M. J. (2011). *IBM SPSS statistics 19 procedures companion.* Reading, MA, Addison-Wesley.

Okazaki, S. (2007). Lessons learned from i-mode: What makes consumers click wireless banner ads? *Computers in Human Behavior, 23*, 1692–1719.

Polymeros, K., Kaimakoudi, E., Schinaraki, M., & Batzios, C. (2015). Analysing consumers' perceived differences in wild and farmed fish. *British Food Journals, 117*, 1007–1016.

Rompre, P. H., Daigle-Landry, D., Guitard, F., Montplaisir, J. Y., & Lavigne, G. J. (2007). Identification of a sleep bruxism subgroup with a higher risk of pain. *Journal of Dental Research, 86*, 837–842.

Rundle-Thiele, S., Kubacki, K., Tkaczynski, A., & Parkison, J. (2015). Using two-step cluster analysis to identify homogeneous physical activity groups. *Marketing Intelligence & Planning, 33*, 522–537.

Stranak, Z., Semberova, J., Barrington, K., O'Donnell, C., Marlow, N., Naulaers, G., et al. (2014). International survey on diagnosis and management of hypotension in extremely preterm babies. *European Journal of Pediatrics, 173,* 793–798.

Tkaczynski, A., & Prebensen, N. K. (2012). French nature-based tourist potentials to Norway: Who are they? *Tourism Analysis, 18*, 181–193.

Tkaczynski, A., Rundle-Thiele, S., & Beaumont, N. (2010). Destination segmentation: A recommended two-step approach. *Journal of Travel Research, 49*, 139–152.

Tkaczynski, A., Rundle-Thiele, S. R., & Prebensen, N. K. (2015). Segmenting potential nature-based tourists based on temporal factors: The case of Norway. *Journal of Travel Research, 54*, 251–265.

Ulstein, I., Wyller, T. B., & Engedal, K. (2007). High score on the relative stress scale, a marker of possible psychiatric disorder in family carers of patients with dementia. *International Journal of Geriatric Psychiatry, 22*, 195–202.